ИНФОРМАТИКА

УДК 519.862.6 DOI 10.25730/VSU.0536.24.002

Эконометрический анализ торговой статистики сети магазинов Rossmann

Шатров Анатолий Викторович[1], Левин Михаил Наумович[2]

[1]доктор физико-математических наук, главный научный сотрудник кафедры ЭВМ,
Вятский государственный университет. Россия, г. Киров; профессор физико-механического института,
Санкт-Петербургский политехнический университет. ORCID: 000-0002-5295-571X. E-mail: shatrov@vyatsu.ru
[2]кандидат физико-математических наук, доцент кафедры прикладной математики и информатики,
Вятский государственный университет. Россия, г. Киров. E-mail: usr00227@vyatsu.ru

Аннотация. В представленной работе выполнена предварительная обработка данных статистической базы продаж магазинов европейской сети Rossmann. В качестве процедуры предварительного анализа используются эконометрические методы обработки данных с платформы Kaggle Rossmann Store Sales. В качестве инструментальных методов используется среда разработки Python. Задача данной статьи заключается в проведении предварительной обработки данных эконометрическими методами. Проанализирована база данных продаж по типам магазинов, временным интервалам работы сети, состоянию спроса потребителей в зависимости от различных факторов. Выполнены корреляционный и дисперсионный анализы статистических данных. Исследованы свойства временных рядов основных факторов, в том числе по факту наличия автокорреляции. Получены распределения продаж как по отдельным типам магазинов, так и всей совокупности сети. Результаты эконометрического анализа необходимы для построения прогностических моделей.

Ключевые слова: математическая статистика, эконометрический анализ, среда разработки Python.

Введение. В качестве предметной области для эконометрического анализа используется база данных торговой статистики Rossmann Store Sales. Эти данные были представлены для международных соревнований по Data Mining (анализ данных) специальным online ресурсом Kaggle (https://www.kaggle.com/), который организует и проводит исследования в области статистического моделирования. Эффективность и достоверность решений определяют независимые эксперты в форме рейтинга решений участников соревнования. Данные представляют заинтересованные фирмы и организации. В представленной статье использованы статистические показатели торговой сети Rossmann (https://www.rossmann.de). Статистические данные продаж сети, представляющие основную (генеральную) совокупность для дальнейшей эконометрической обработки, разделены на две выборки: обучающую (train.csv) и тестовую (test.csv), находящиеся в соотношении 4:1.

Кроме этого, сформирована вспомогательная выборка, представленная файлом store.csv, содержащим дополнительную информацию об условиях работы торговой сети. Задача предварительной обработки данных в рамках эконометрического анализа [1; 2] данных статистики продаж состоит в структурировании данных, определении закономерностей между факторами, формировании новых переменных из комбинаций исходных факторов, исследовании динамики временных рядов [3].

Предварительная обработка и анализ данных сети магазинов. Прежде чем переходить к обработке статистики, требуется провести предварительный анализ данных. Для этого в среде разработки Python предназначена библиотека Pandas [8]. Структура обучающей выборки представлена на рис 1.

Генеральная совокупность представляет матрицу из 1017209 строк и 8 столбцов. В столбцах перечислены факторы (значения этих факторов на указанную дату):

– Store – порядковый номер магазина;

– Day_Of_Week – количество рабочих дней в текущей неделе;

– Sales – объем продаж на указанную дату (целевая переменная);

– Customers – количество клиентов на указанную дату;

– Open – статус работы магазина: 0 – закрыт, 1 – открыт;

– Promo – статус проведения промоакции: 0 – нет, 1 – да;

– State_Holiday – указывает на государственный праздник;

– School_Holiday – показатель наличия школьных каникул: 0 – нет, 1 – да.

В строках указаны даты наблюдения перечисленных факторов.

Date	Store	DayOfWeek	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
2015-04-24	731	5	6498	751	1	0	0	0
2015-04-24	732	5	4187	484	1	0	0	0
2015-04-24	733	5	14836	3767	1	0	0	0
2015-04-24	734	5	3935	469	1	0	0	0
2015-04-24	735	5	3987	446	1	0	0	0
2015-04-24	736	5	3763	408	1	0	0	0
2015-04-24	737	5	4788	739	1	0	0	0

Рис. 1. Фрагмент обучающей выборки train.csv

В целях структурирования данных вводится новая переменная Sale_Per_Customer (средний чек) – отношение общего объема продаж к количеству клиентов на указанную дату. Если продажи отсутствуют, средний чек равен 0. На рис. 2 показано, как изменялись продажи по месяцам за период наблюдений с 2013.01 по 2015.06. График изменения фактора Sales демонстрирует рост среднемесячного объема продаж. Линейный тренд зависимости фактора Sales от времени является положительным и представлен уравнением регрессии $Sales = 15.725t + 5531.4$. На рис. 3 представлены диаграммы среднегодовых значений факторов Sales и Customers.

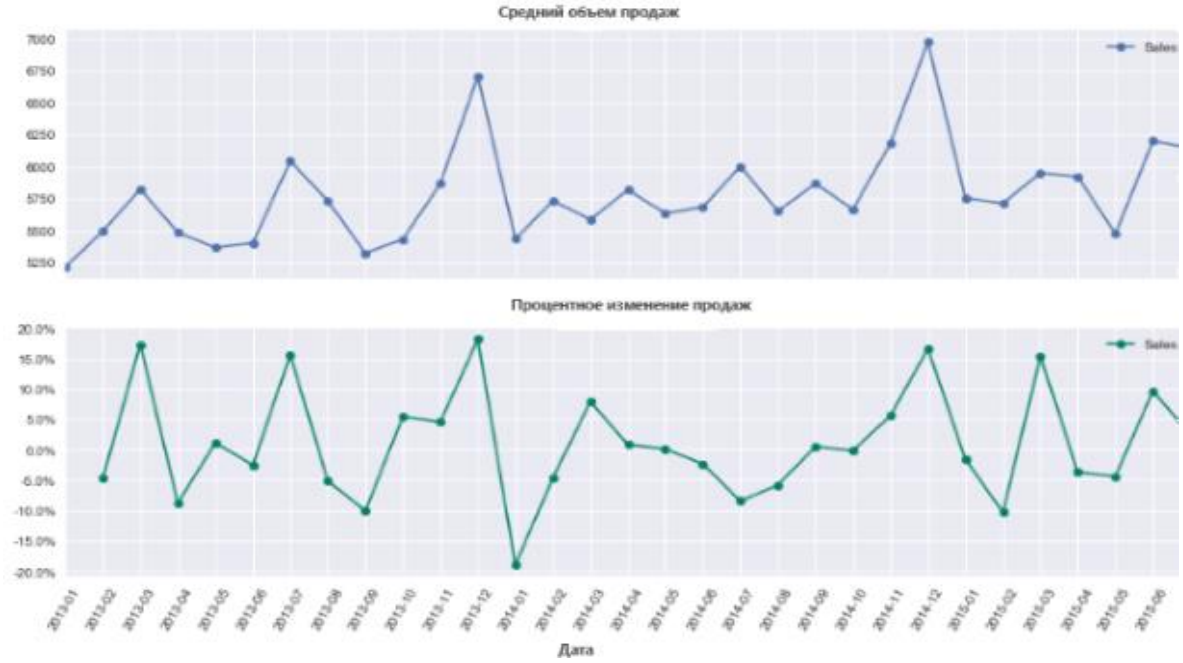

Рис. 2. График помесячного изменения фактора Sales

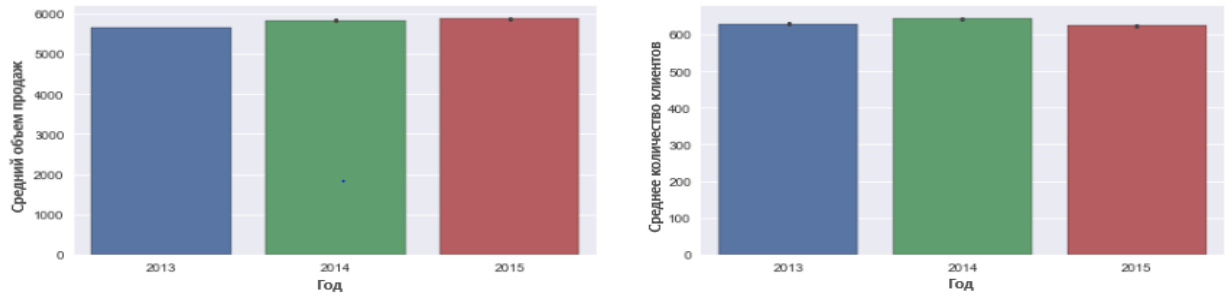

Рис. 3. Средний объем продаж и среднее количество клиентов в год

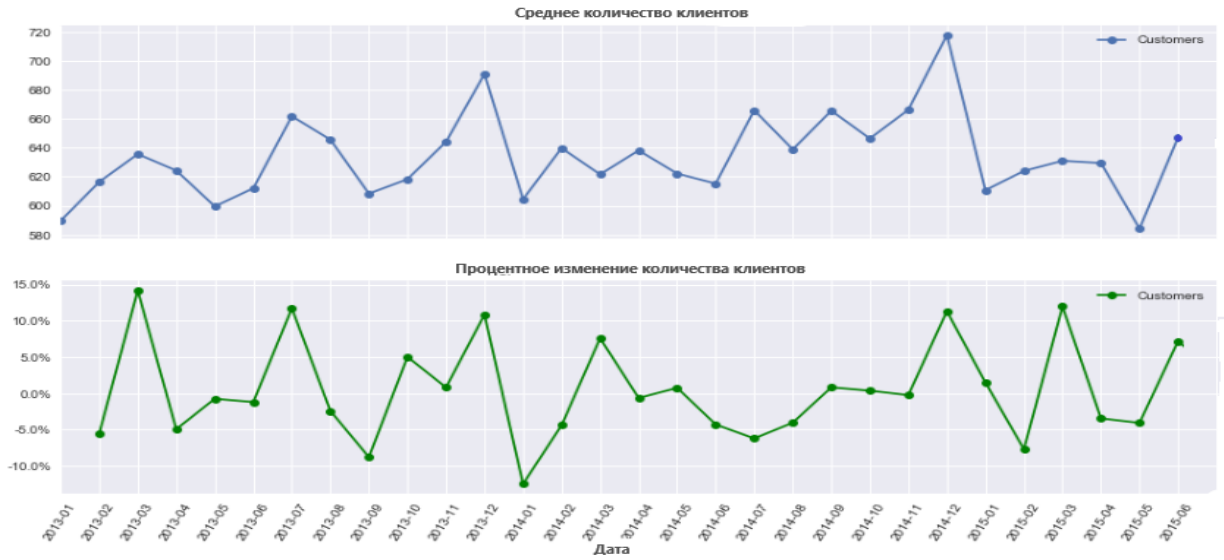

Рис. 4. График изменения фактора Customers

На рис. 4 показано изменение количества клиентов за весь период по месяцам, как это было сделано ранее для объема продаж.

Функция ECDF (ECDF – empirical cumulative distribution function) из пакета stats-models.distributions [7; 11] дает представление о непрерывном распределении для факторов Sales, Customers, Sale_Per_Customer.

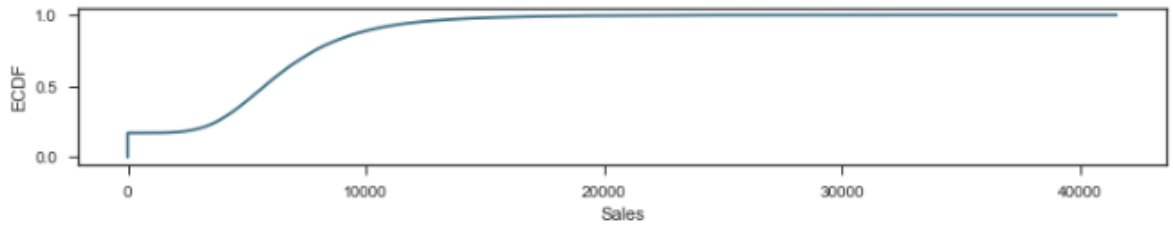

Рис. 5. Эмпирическое распределение по фактору Sales

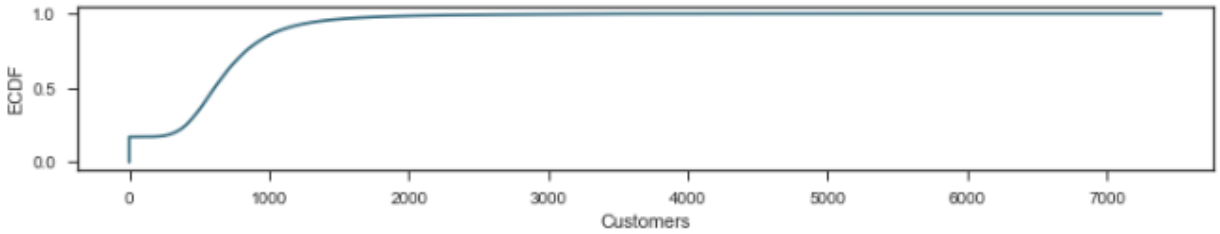

Рис. 6. Эмпирическое распределение по фактору Customers

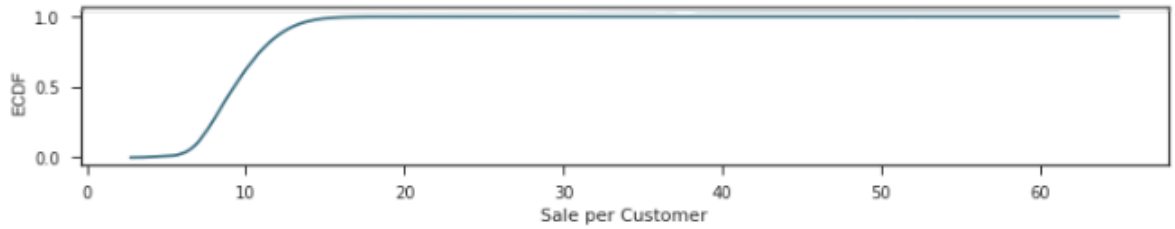

Рис. 7. Эмпирическое распределение по фактору Sale_Per_Customer

Эти графики дают информацию в виде оценки порядков изменения факторов. Так, например, около 20 % наблюдений за фактором продаж показывают нулевые значения фактора Sales (рис. 5). Около 20 % наблюдений количества клиентов демонстрируют нулевые значения фактора Customers (рис. 6). Это означает, что структурирование выборки по этим факторам необходимо производить с учетом влияния нулевых значений продаж и количества клиентов. Учет этого влия-

ния заключается в удалении строк с нулевыми значениями этих факторов. На примере фактора Sales построим график распределения ежедневных продаж для всех магазинов после удаления нулевых наблюдений (рис. 8).

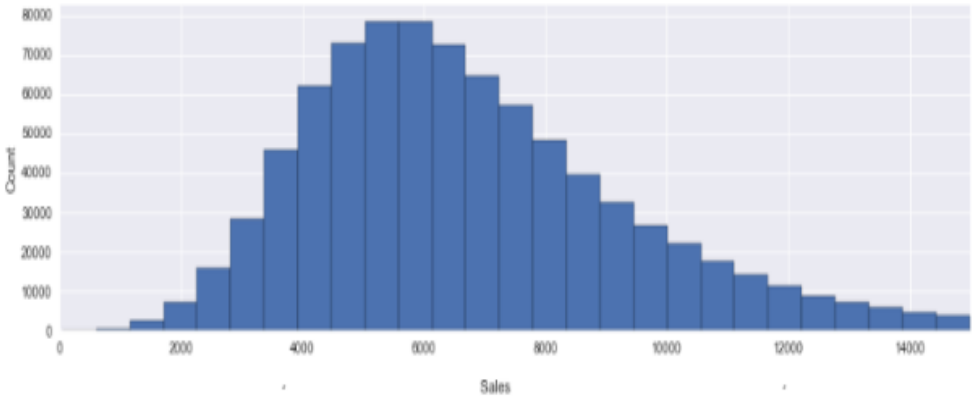

Рис. 8. Распределение всех типов магазинов по фактору Sales

Влияние дополнительных факторов на итоги продаж. Существенное влияние на структуру генеральной совокупности оказывает вспомогательная выборка store.csv. Эта выборка позволяет оценить влияние дополнительных факторов на результаты деятельности торговой сети с точки зрения эффективности продаж и привлечения клиентов. Структура файла store.csv включает в себя следующие факторы:

– Store – идентификатор магазина;

– Store_Type – типы магазинов: A, B, C, D;

– Assortment – признаки ассортимента: a – базовый, b – дополнительный, c – расширенный;

– Competition_Distance – расстояние в метрах до ближайшего магазина конкурентов;

– Competition_Open_Since[Month/Year] – дата открытия магазина ближайшего конкурента;

– Promo2 – факт участия магазина в дополнительной рекламной акции: 0 – не участвует, 1 – участвует;

– Promo2_Since[Year/Week] – дата начала участия магазина в Promo2;

– Promo_Interval – интервалы (раунды) проведения магазинами акций Promo2.

Учитывая дополнительную выборку store.csv, можно проанализировать иерархию магазинов по объемам продаж с учетом их типа. Рассмотрим различные уровни фактора Store_Type и соответствующие им показатели и основные характеристики фактора Sales.

Таблица 1

Характеристики фактора Store_Type по объему продаж

Тип магазина	Кол-во	Ср. знач.	Стд. откл. (RSME)	min	25 %	50 %	75 %	max
A	457042	6925.70	3277.35	46	4696.25	6285.00	8406.00	41551
B	15560	10233.38	5155.73	1252	6345.75	9130.00	13184.25	38722
C	112968	6933.13	2896.96	133	4916.00	6408.00	8349.25	31448
D	258768	6822.30	2556.40	538	5050.00	6395.00	8123.25	38037

Результаты работы магазинов по объему продаж за весь период показывают, что магазины типа B имеют самое высокое среднее значение продаж. Распределение долей продаж по квартилям для этого типа по сравнению с другими менее однородно, что отражает относительно большое значение RSME. При этом наибольшая часть объема продаж магазинов типа B принадлежит третьему квартилю.

Таблица 2

Общий объем продаж и количество клиентов по каждому типу магазинов за весь период

Тип магазина	Количество клиентов	Объем продаж
A	**363541431**	**3165334859**
B	31465616	159231395
C	92129705	783221426
D	156904995	1765392943

Результаты работы магазинов по объему продаж и количеству клиентов (факторы Sales и Customers) представлены в таблице 2. Данные этой таблицы демонстрируют очевидную иерархию по суммарному показателю «Sales + Customers»: $A \succ D \succ C \succ B$ (здесь знак + означает объединение факторов, знак $\succ$ соответствует предпочтительности факторов Store_Type по суммарному показателю).

Ниже приведены графики динамики факторов Sales и Customers без учета рекламной акции (фактор Promo=0) и с учетом рекламной акции (Promo=1).

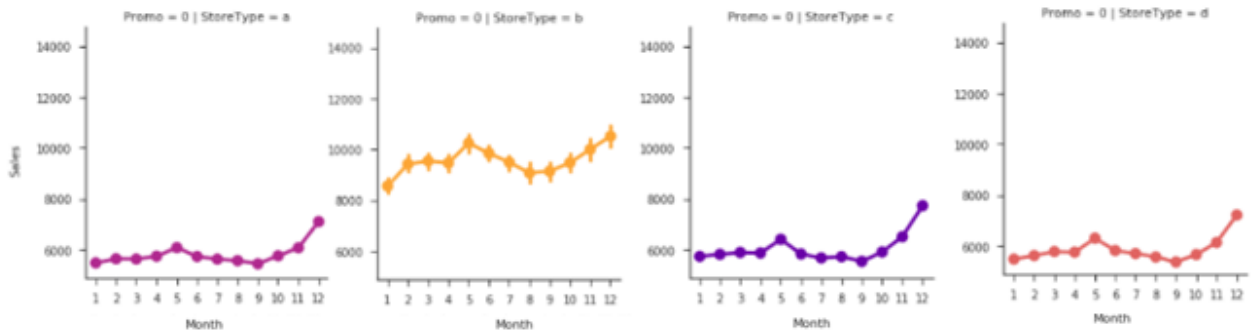

Рис. 9. Динамика объема продаж по каждому типу магазинов (A, B, C, D – слева направо) без учета рекламной акции по месяцам (Promo=0)

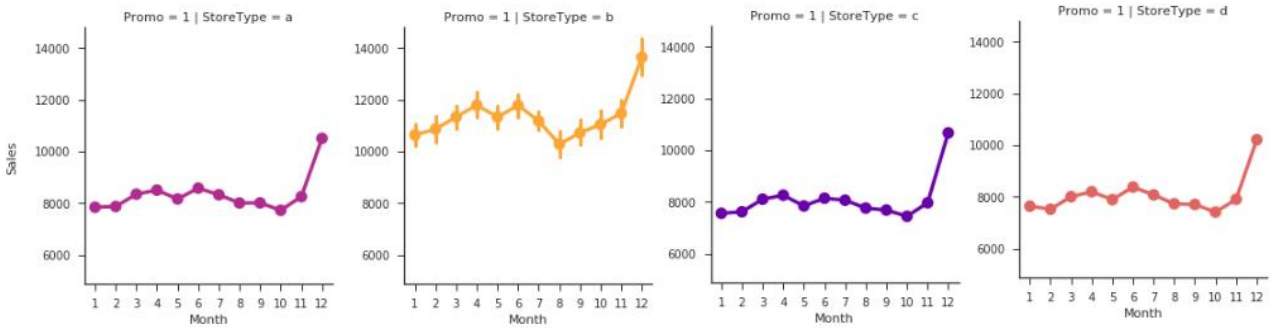

Рис. 10. Динамика объема продаж по каждому типу магазинов (A, B, C, D – слева направо) с учетом рекламной акции по месяцам (Promo=1)

Графики динамики фактора Sale_Per_Customer (среднего чека на одного клиента) представлены в последовательности факторов Store_Type (типы магазинов A, B, C, D – следуют слева направо). Наиболее эффективными по фактору Sale_Per_Customer оказываются магазины типа D: среднее значение фактора 12.26€ (Promo=1) и 10.63€ (Promo=0) по наблюдениям в течение года. Наименее эффективными по фактору Sale_Per_Customer оказываются магазины типа B: среднее значение фактора 5.58€ (Promo=1) и 5.26€ (Promo=0) по наблюдениям в течение года.

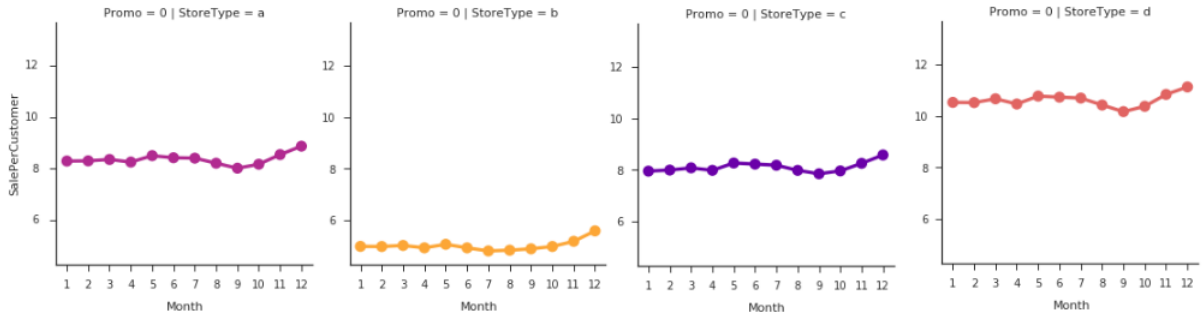

Рис. 11. Динамика среднего чека на одного клиента по каждому типу магазинов без учета рекламной акции по месяцам (Promo=0)

Низкое значение фактора Sale_Per_Customer для магазинов типа B объясняет тот факт, что несмотря на объемы продаж (по данным таблицы 1, рис. 9 и рис. 10, показатели фактора Sales являются наибольшими), магазины этого типа не являются эффективными.

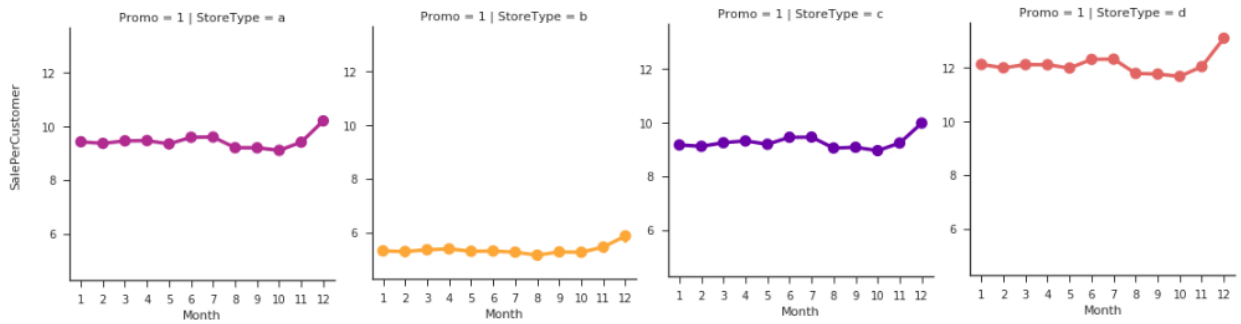

Рис. 12. Динамика среднего чека на одного клиента по каждому типу магазинов с учетом рекламной акции по месяцам (Promo=1)

При этом магазины типа B занимают специфическую нишу: в них клиенты совершают много покупок товаров из низкого ценового ассортимента.

Оценку влияния фактора Competition_Distance на статистику продаж можно получить по диаграмме рассеяния фактора Sales. Роль фактора Competition_Distance заключается в определении расстояния от фиксированного магазина до ближайшего конкурента. На рис. 13 приведен график корреляционного поля зависимости объема продаж от расстояния до ближайшего конкурента. График показывает отрицательную корреляционную связь между факторами Competition_Distance и Sales.

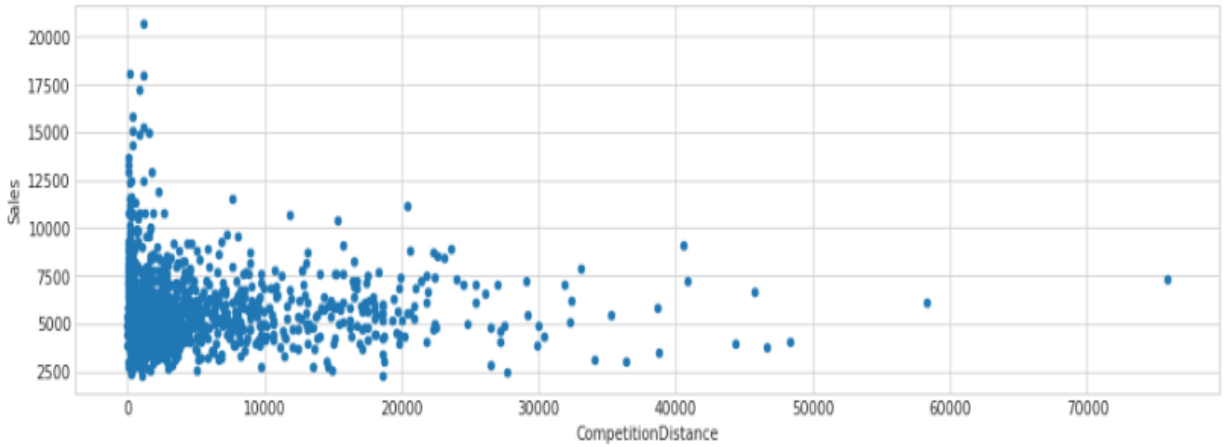

Рис. 13. Диаграмма рассеяния зависимости объема продаж и расстояния до ближайшего конкурента (фактор Competition_Distance)

При малых значениях фактора Competition_Distance разброс суммарных продаж является наибольшим и колеблется в области сгущения от 2000€ до 7500 €, при больших значениях фактора Competition_Distance разброс существенно уменьшается и устанавливается в среднем около 5000€. Обучающую выборку train.csv после присоединения к ней дополнительной выборки store.csv используем для анализа корреляционной зависимости между факторами объединенной выборки. Тесноту связи между факторами можно оценить по тепловой карте корреляций [10] (рис. 14). Графическое представление корреляционных связей между факторами дает исчерпывающую информацию для анализа факторного взаимодействия в объединенной выборке. Так, можно видеть, что фактор Customers показывает сильную положительную корреляцию с объемом продаж – коэффициент корреляции выше 0.8. Наблюдается положительная корреляция между факторами Promo и Sales. На рис. 15 представлен график динамики продаж для магазинов A, B, C и D. Следует отметить, что прослеживается увеличение объема продаж для A, B и D, но не для магазинов типа C (третий график сверху).

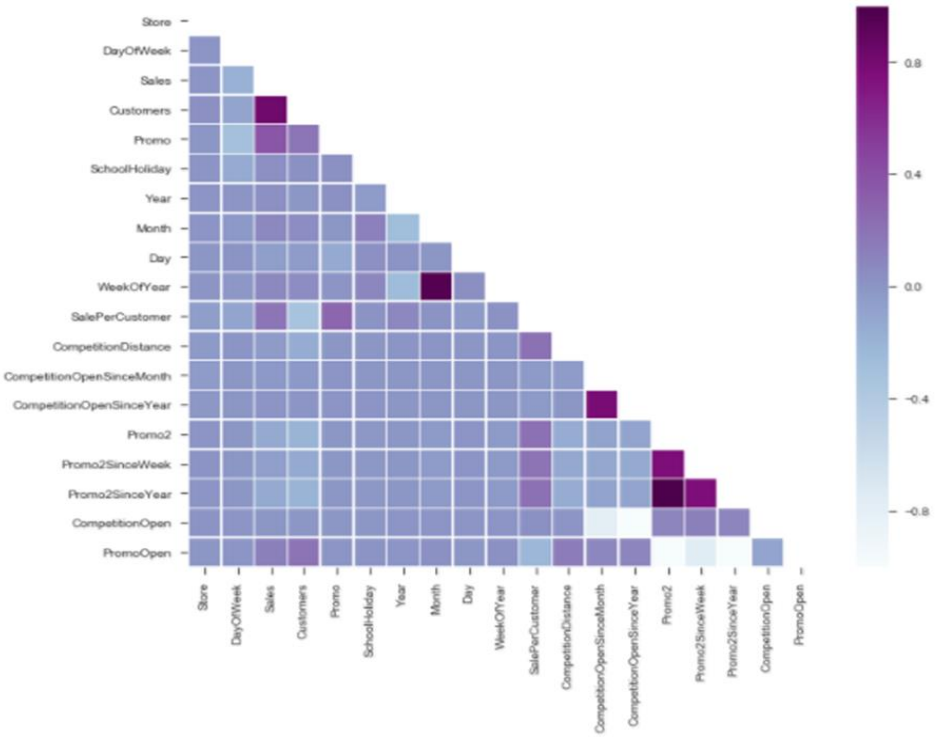

Рис. 14. Тепловая карта общих корреляций

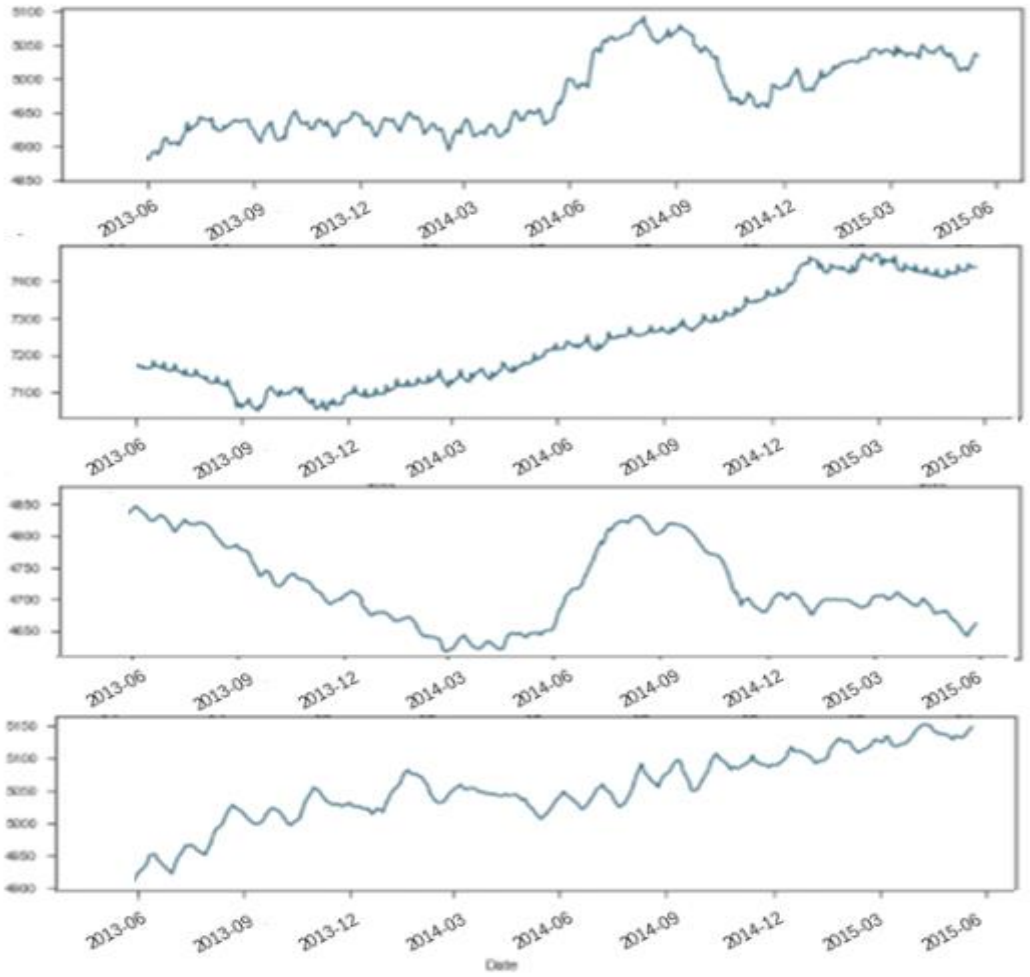

Рис. 15. Динамика продаж магазинов по типам A, B, C и D

Проверка стационарности временных рядов. Проверка стационарности временного ряда осуществляется по коэффициентам выбранной регрессионной модели или с помощью числовых характеристик временного ряда [1–3]. Наиболее адекватной оценкой явлений нестационарности временных рядов являются автокорреляционные функции (АКФ и ЧАКФ). Ниже на рис. 16–19 представлены графики АКФ и ЧАКФ для каждого типа магазинов.

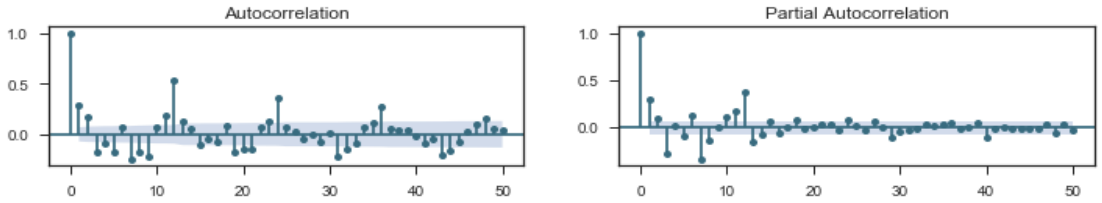

Рис. 16. Автокорреляционная и частная автокорреляционная функции для магазинов типа A

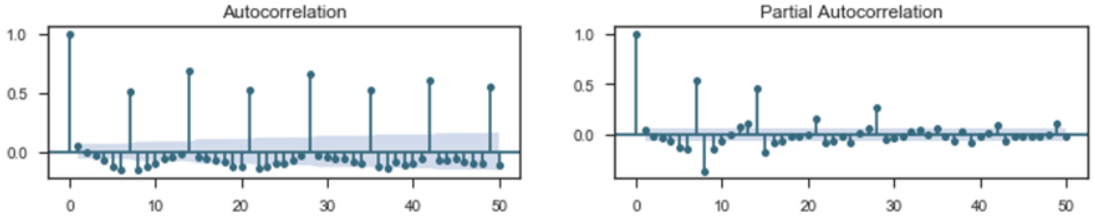

Рис. 17. Автокорреляционная и частная автокорреляционная функции для магазинов типа B

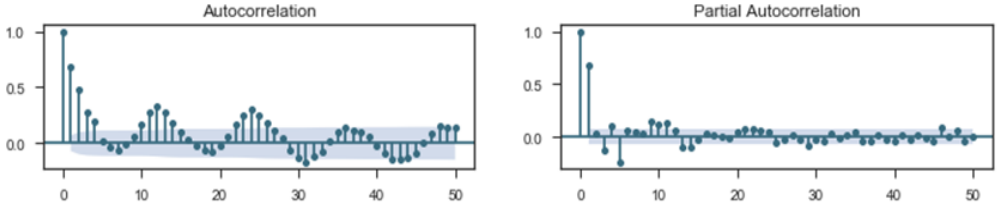

Рис. 18. Автокорреляционная и частная автокорреляционная функции для магазинов типа C

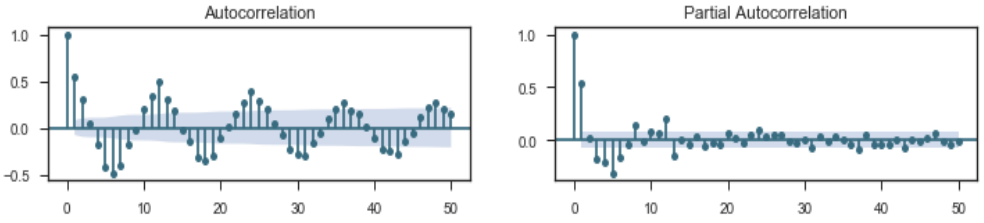

Рис. 19. Автокорреляционная и частная автокорреляционная функции для магазинов типа D

Для магазинов типа A и типа B можно сделать следующий вывод: оба типа показывают сезонности при определенных задержках. Для типа A каждое двенадцатое наблюдение определяется положительными пиками в 12 (s) и 24 (2s) лагах и так далее. Для типа B это недельный тренд с положительными всплесками в 7 (s), 14 (2s), 21 (3s) и 28 (4s) лагах. В целом можно сделать вывод, что временные ряды по основным факторам являются стационарными. Анализ автокорреляций временных рядов (ЧАКФ основных факторов убывают) показывает, что выполняется предположение о гомоскедастичности, то есть дисперсия временных рядов является практически однородной.

Заключение. В данной работе представлены результаты предварительной обработки данных статистики продаж крупной европейской компании Rossmann, база данных которой была представлена на сайте международного ресурса по анализу данных Kaggle. Необходимость эконометрической предварительной обработки данных заключается в подготовке статистических выборок к последующему интеллектуальному анализу. На данном этапе выявлена целевая функция Sales, определяющая объемы продаж, выполнено структурирование базы данных по выявлению основных факторов, определяющих изменение целевой функции. Произведена оценка влияния вы-

деленных факторов Customers, Sale_Per_Customer, Store, Store_Type, Promo, Promo2, Competition_Distance на целевую функцию. Анализ временных рядов продаж по факторам Store_Type показал, что они являются стационарными и пригодными для последующего моделирования. Выполненный эконометрический анализ является необходимым этапом для последующего этапа создания моделей прогнозирования динамики поведения целевой функции. Результаты моделирования представлены в следующей статье авторов журнала[1].

Список литературы

1. *Магнус Я. Р., Катышев П. К., Пересецкий А. А.* Эконометрика. Начальный курс. М. : Дело, 2007. 504 с.

2. *Нестеров С. А.* Базы данных. Интеллектуальный анализ данных : учебное пособие. СПб. : Изд-во Политехн. ун-та, 2011. 272 с.

3. *Box G., Jenkins G.* Time series analysis: forecasting and control. John Wiley and Sons, 2008. P. 748.

4. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. 2nd edition. Springer, 2009. 763 p.

5. *Fayyad M., Piatetsky-Shapiro G., Smyth P.* From Data Mining to Knowledge Discovery in Databases. URL: https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf (дата обращения: 01.02.2023).

6. NumPy User Guide (Release 1.14.0). URL: http://docs.scipy.org/doc/numpy/user/ (дата обращения: 11.05.2023).

7. Pandas: powerful Python data analysis toolkit (Release 0.23.0). URL: http://pandas.pydata.org/pandas-docs/stable/ (дата обращения: 11.05.2023).

8. *Piatetsky-Shapiro G.* CRISP-DM, still the top methodology for analytics, data mining, or data science projects. 2014. URL: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-datascience-projects.html (дата обращения: 25.05.2023).

9. Scikit-learn user guide (Release 0.19.1). URL: http://scikit-learn.org/stable/user_guide.html (дата обращения: 11.05.2023).

10. Seaborn: Annotated heatmaps (Release 0.9.0). URL: http://seaborn.pydata.org/examples/heatmap_annotation.html (дата обращения: 11.05.2023).

11. Statsmodels: statistics in Python (Release 0.9.0). URL: http://www.statsmodels.org/stable/index.html (дата обращения: 11.05.2023).

Econometric analysis of trade statistics of the Rossmann chain of stores

Shatrov Anatoly Viktorovich[1], Levin Mikhail Naumovich[2]

[1]Doctor of Physical and Mathematical Sciences, Chief Researcher of the Computer Department, Vyatka State University. Russia, Kirov; professor of the Institute of Physics and Mechanics, St. Petersburg Polytechnic University. Russia, St. Petersburg. ORCID: 0000-0002-5295-571X. E-mail: shatrov@vyatsu.ru

[2]PhD in Physical and Mathematical Sciences, associate professor of the Department of Applied Mathematics and Computer Science, Vyatka State University. Russia, Kirov. E-mail: usr00227@vyatsu.ru

Abstract. In the presented work, preliminary processing of data from the statistical database of sales of stores of the European Rossmann chain has been performed. Econometric methods of data processing from the Kaggle Rossmann Store Sales platform are used as a preliminary analysis procedure. The Python development environment is used as instrumental methods. The purpose of this article is to carry out preliminary data processing using econometric methods. The sales database was analyzed by types of stores, time intervals of the network, the state of consumer demand, depending on various factors. Correlation and variance analyses of statistical data were performed. The properties of the time series of the main factors are investigated, including the presence of autocorrelation. Sales distributions are obtained both for individual types of stores and for the entire network. The results of the econometric analysis are necessary for the construction of predictive models.

Keywords: mathematical statistics, econometric analysis, Python development environment.

References

1. *Magnus Ya. R., Katyshev P. K., Pereseckij A. A. Ekonometrika. Nachal'nyj kurs* [Econometrics. The beginners' course]. M. Delo (Business), 2007. 504 p.

2. *Nesterov S. A. Bazy dannyh. Intellektual'nyj analiz dannyh : uchebnoe posobie* [Databases. Data mining : textbook]. SPb. Publishing House of the Polytechnic University, 2011. 272 p.

3. *Box G., Jenkins G.* Time series analysis: forecasting and control. John Wiley and Sons, 2008. P. 748.

4. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. 2nd ed. Springer, 2009. 763 p.

5. *Fayyad M., Piatetsky-Shapiro G., Smyth P.* From Data Mining to Knowledge Discovery in Databases. Available at: https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf (date accessed: 01.02.2023).

[1] Статья авторов «Сравнение двух методов анализа данных по продажам в сети магазинов Rossmann» в следующем выпуске журнала.

6. NumPy User Guide (Release 1.14.0). Available at: http://docs.scipy.org/doc/numpy/user/ (date accessed: 11.05.2023).

7. Pandas: powerful Python data analysis toolkit (Release 0.23.0). Available at: http://pandas.pydata.org/pandas-docs/stable/ (date accessed: 11.05.2023).

8. *Piatetsky-Shapiro G.* CRISP-DM, still the top methodology for analytics, data mining, or data science projects. 2014. Available at: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-datascience-projects.html (date accessed: 25.05.2023).

9. Scikit-learn user guide (Release 0.19.1). Available at: http://scikit-learn.org/stable/user_guide.html (date accessed: 11.05.2023).

10. Seaborn: Annotated heatmaps (Release 0.9.0). Available at: http://seaborn.pydata.org/examples/heatmap_annotation.html (date accessed: 11.05.2023).

11. Statsmodels: statistics in Python (Release 0.9.0). Available at: http://www.statsmodels.org/stable/index.html (date accessed: 11.05.2023).