Разведывательный анализ данных

Цель отчета: первоначальный осмотр и обработка данных для дальнейших исследований,

Изучение самых простых взаимосвязей между полями таблиц.

Данные: https://www.kaggle.com/c/rossmann-store-sales/data

Оглавление

Изучение общего вида данных полей Sales, Customers, CompetitionDistance………………………………………2-3

Изучение корреляции между основными полями……………………………………………………………………………………4

Изучение влияния поля SchoolHoliday на объем продаж…………………………………………………………………………5

Дополнение…………………………………………………………………………………………………………………………………………………5

Цель отчета: первоначальный осмотр и обработка данных для дальнейших исследований,

Изучение самых простых взаимосвязей между полями таблиц.

Данные: https://www.kaggle.com/c/rossmann-store-sales/data

Общий вид данных

Данные в полей ‘Sale’ и ‘Customers’ таблицы ‘Train’ имеют нормальный вид, с большим “хвостом ” данных в правой части графика

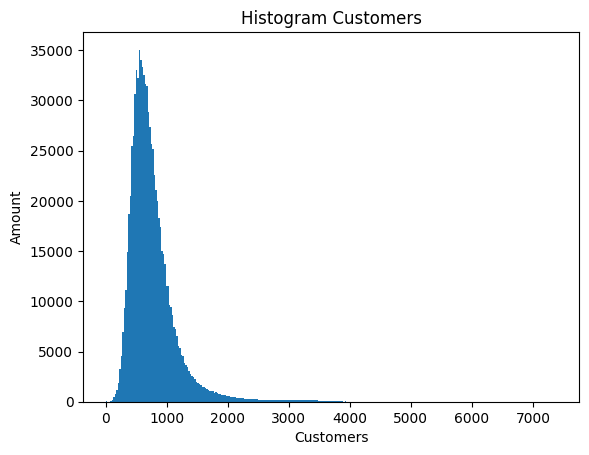
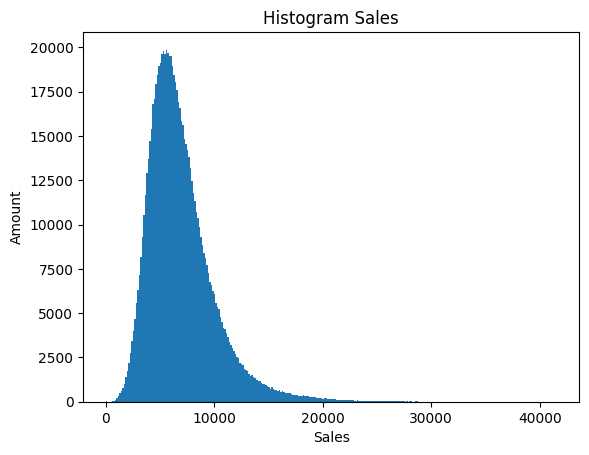


Рис 1. Рис2.

‘Ящики’ Графиков «Boxplot»(находятся в приложении) имеют сжатый вид, а также наблюдается большое значение выбросов, что может говорить о большом количестве магазинов, имеющих показатели, сильно отличающиеся от нормальных по выборке.

Данные поля ‘CompetitionDistance’ таблицы ‘Store’ имеют вид не нормального распределения, как и у предыдущих графиков - можно отметить большой “хвост” данных в правой части графика. Из этого можно сделать вывод, что большая часть магазинов имеет малую дистанцию и находятся в зоне высокой конкуренции.

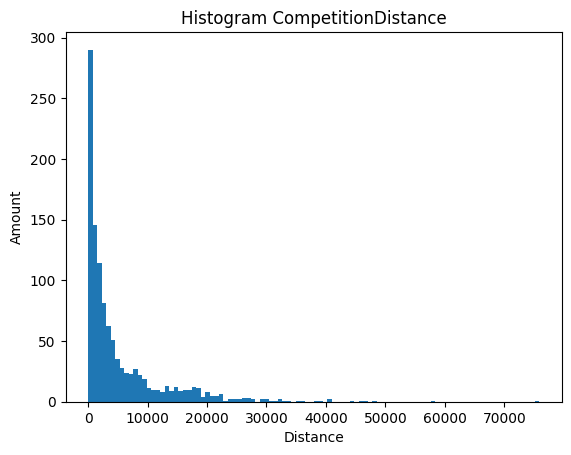


Рис3.

Далее проведены исследования для всех магазинов и сумм их продаж, отсортированных по убыванию суммы продаж.

В таблице 1 можно увидеть зависимость процента всех магазинов на процент всех продаж.

На Рисунке 4 изображен график продаж и магазинов.

|  |  |
| --- | --- |
| Магазин, % | Продажи, % |
| ≈5 | 10 |
| ≈11 | 20 |
| ≈19 | 30 |
| ≈28 | 40 |
| ≈37 | 50 |
| ≈47 | 60 |
| ≈58 | 70 |
| ≈70 | 80 |
| ≈83 | 90 |

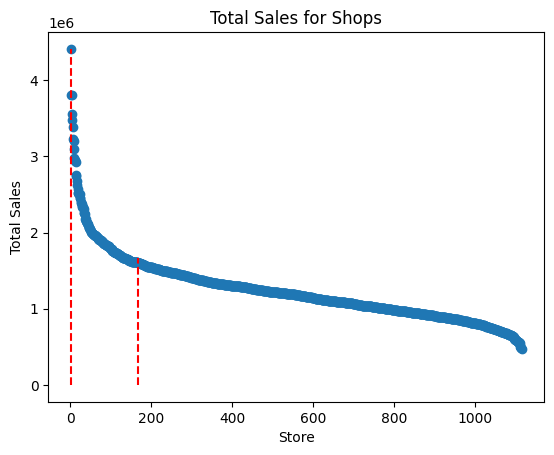


Рис4.

Линия продаж на большей части графика - имеет достаточно ровный и прямой вид, из чего мы можем сделать вывод о равном распределении прибыли и отсутствии рисков крупных финансовых потерь, при закрытии нескольких случайных магазинов.

Так же можно отметить особо прибыльные магазины, а именно 11%, которые обеспечивают 20% всех продаж, они представлены на графике [номер графика], между красными линиями.

Корреляция между полями

‘Sales-Customers’ и 'Customers-CompetitionDistance'

Для численных полей обеих таблиц было проведено исследование корреляции, по итогам которых, было выявлено:

Поля ‘Sales’ и ‘Customers’ имеют положительную корреляцию 0.82 что может говорить о хорошем качестве товара и цене, благодаря чему 80 процентов посетителей совершает покупку

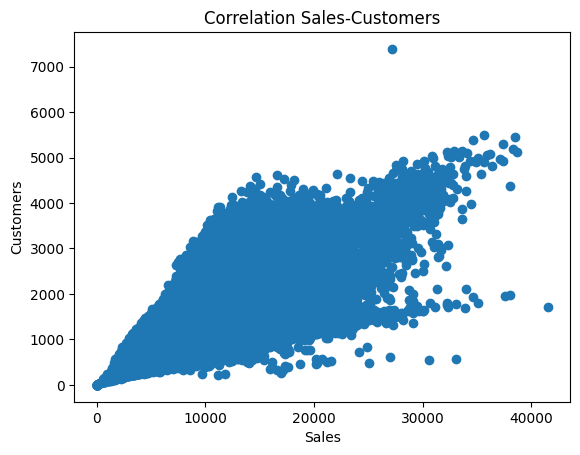


Рис5.

Так же было проведено исследование для полей 'Sales' и ‘CompetitionDistance’, которое выявило отсутствие корреляции 0.003, что может говорить о лояльности покупателя к бренду магазина и игнорировании конкурентов, даже на небольших расстояниях.

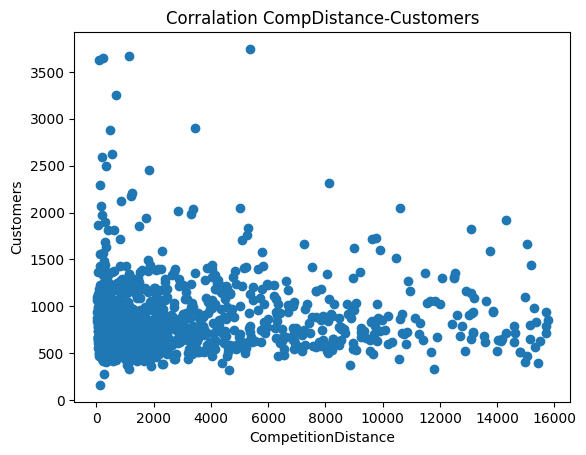


Рис6.

Влияние ‘SchoolHoliday’ на объем продаж

Для исследования были взяты магазины ‘StoreType’ = ‘a’, ‘AssortmentType’ = ‘a’, так же для точности были взяты только магазины не имеющие конкурентов и не участвующие в промо-акциях.

Итоги исследования показали падение общей суммы продаж на 76% и прирост средних дневных продаж на 8%

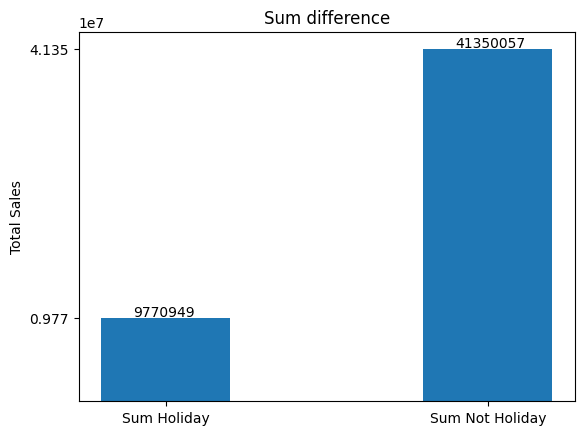
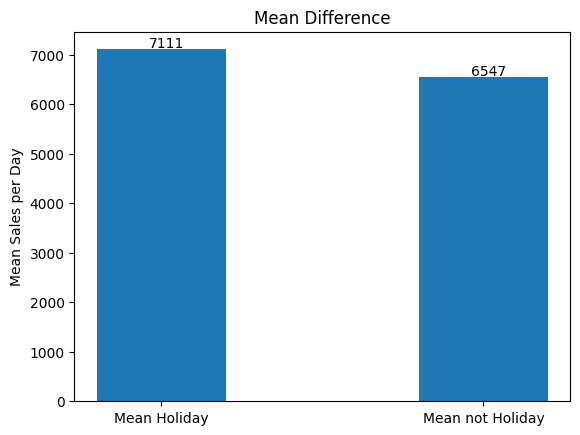
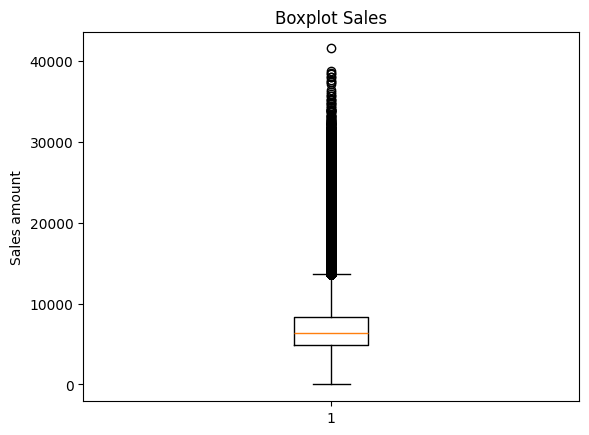
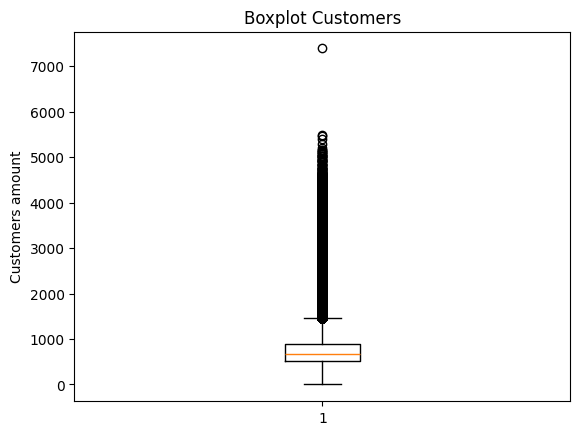
 

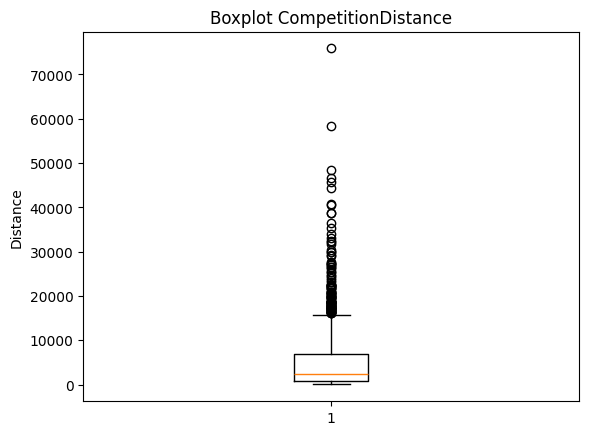
Рис7. Рис8.

По итогам исследования можно сказать, что поле «SchoolHoliday» в целом негативно влияет на выручку, за счет отстутсвия регулярного трафика от учащихся, однако поднимает средние продажи, благодоря более масштабным покупкам от старшего поколения.

Дополнение



[519.0, 676.0, 893.0] [4859.0, 6369.0, 8360.0]



[712.5, 2325.0, 6887.5]