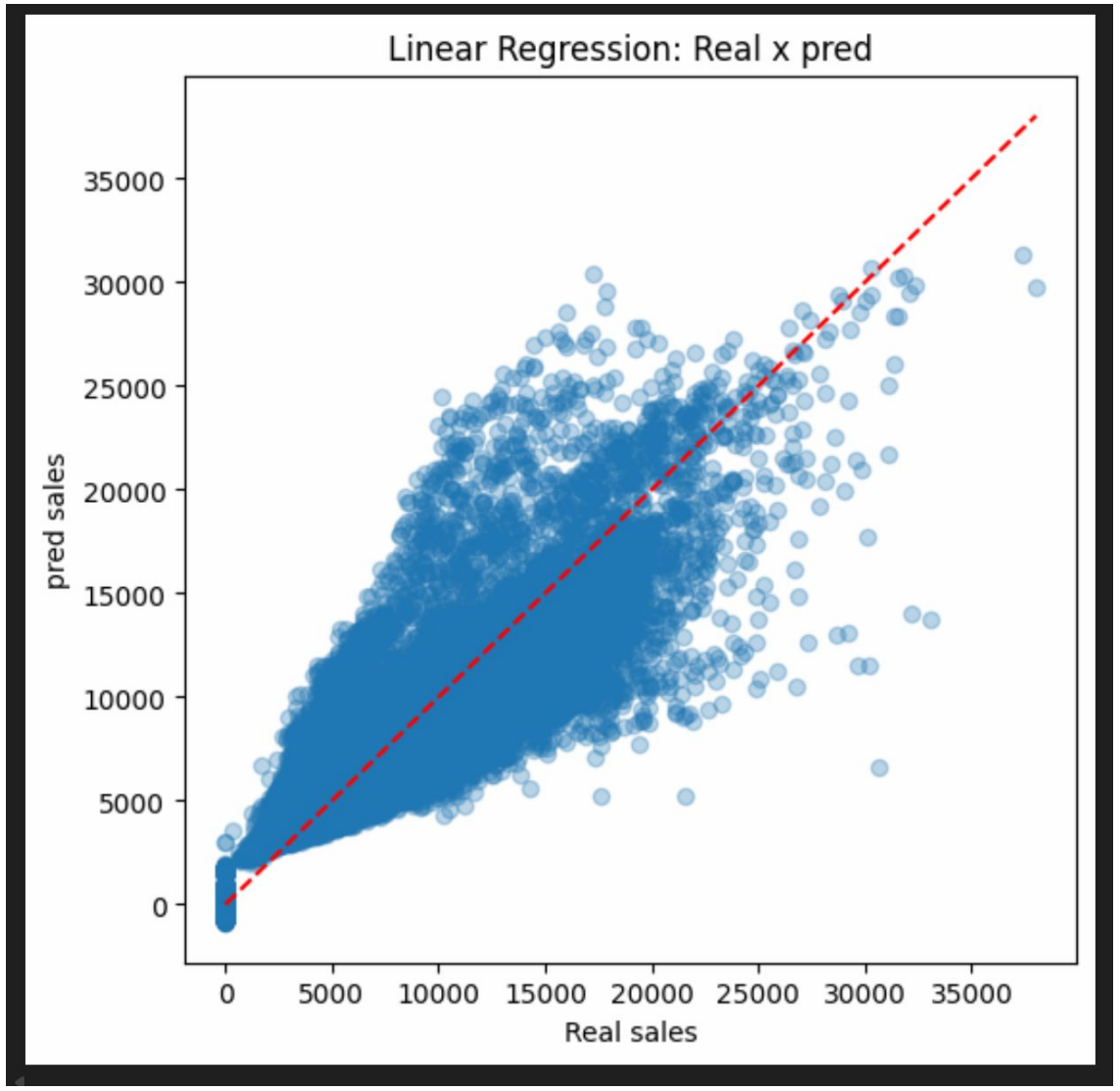
Sales Predictions

Group name : G1
Ahmad
Duda
Georg
Inna
Tiago

Executive summary

- Accuracy with training data (sales.csv)
- Best model - Linear Regression
- R^2 Prediction - real life data 80%
- Quick recap of alternatives considered
- Other important considerations



Exploring the Data

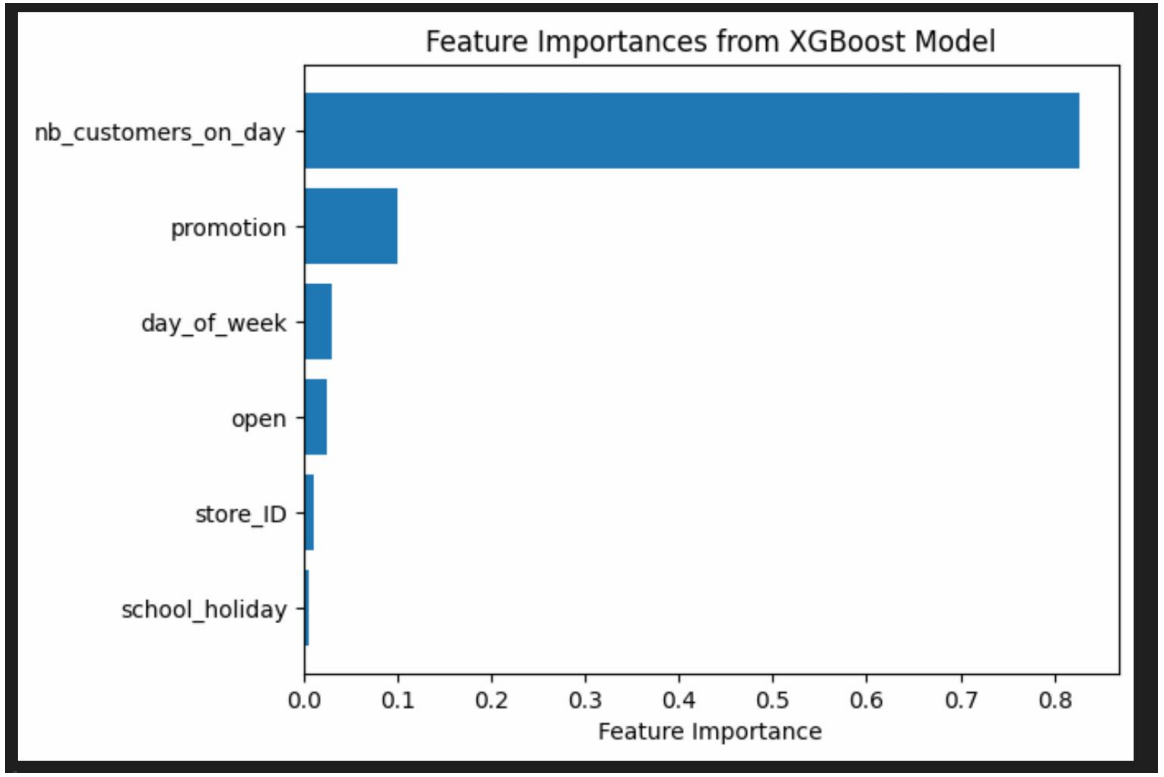
- First, we explored the data with print(df.info())

```
<class 'pandas.core.frame.DataFrame'>
Index: 640840 entries, 425390 to 305711
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   store_ID             640840 non-null  int64
 1   day_of_week          640840 non-null  int64
 2   date                 640840 non-null  object
 3   nb_customers_on_day  640840 non-null  int64
 4   open                 640840 non-null  int64
 5   promotion            640840 non-null  int64
 6   state_holiday        640840 non-null  object
 7   school_holiday       640840 non-null  int64
 8   sales                640840 non-null  int64
dtypes: int64(7), object(2)
memory usage: 48.9+ MB
None
```

-> non numerical data has to be transformed

Data Cleaning

- Checked for feature importance

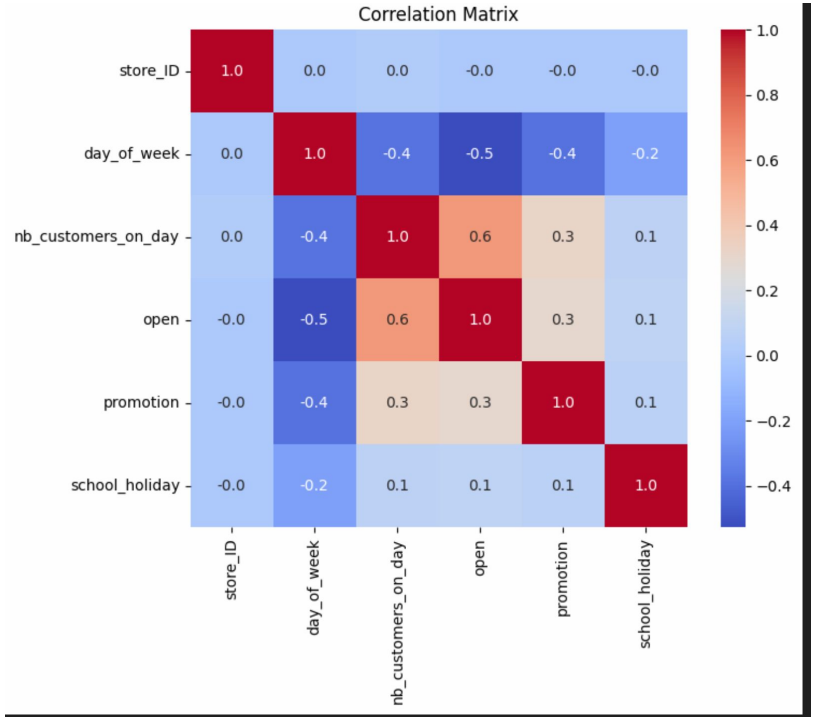


-> Important features: nb_customers_on_day, promotion, day_of_week, open

-> Dropped Unnamed col (or *index*), because while checking it on a neat excel file it looked like a simple column of id's with no relation to the rest of the data (low feature importance).

Data Cleaning

- We checked the correlation between the features.



-> Found correlation between Number of Customers On Day VS Sales.
Sales chosen as *target* for prediction.

Further Data Cleaning

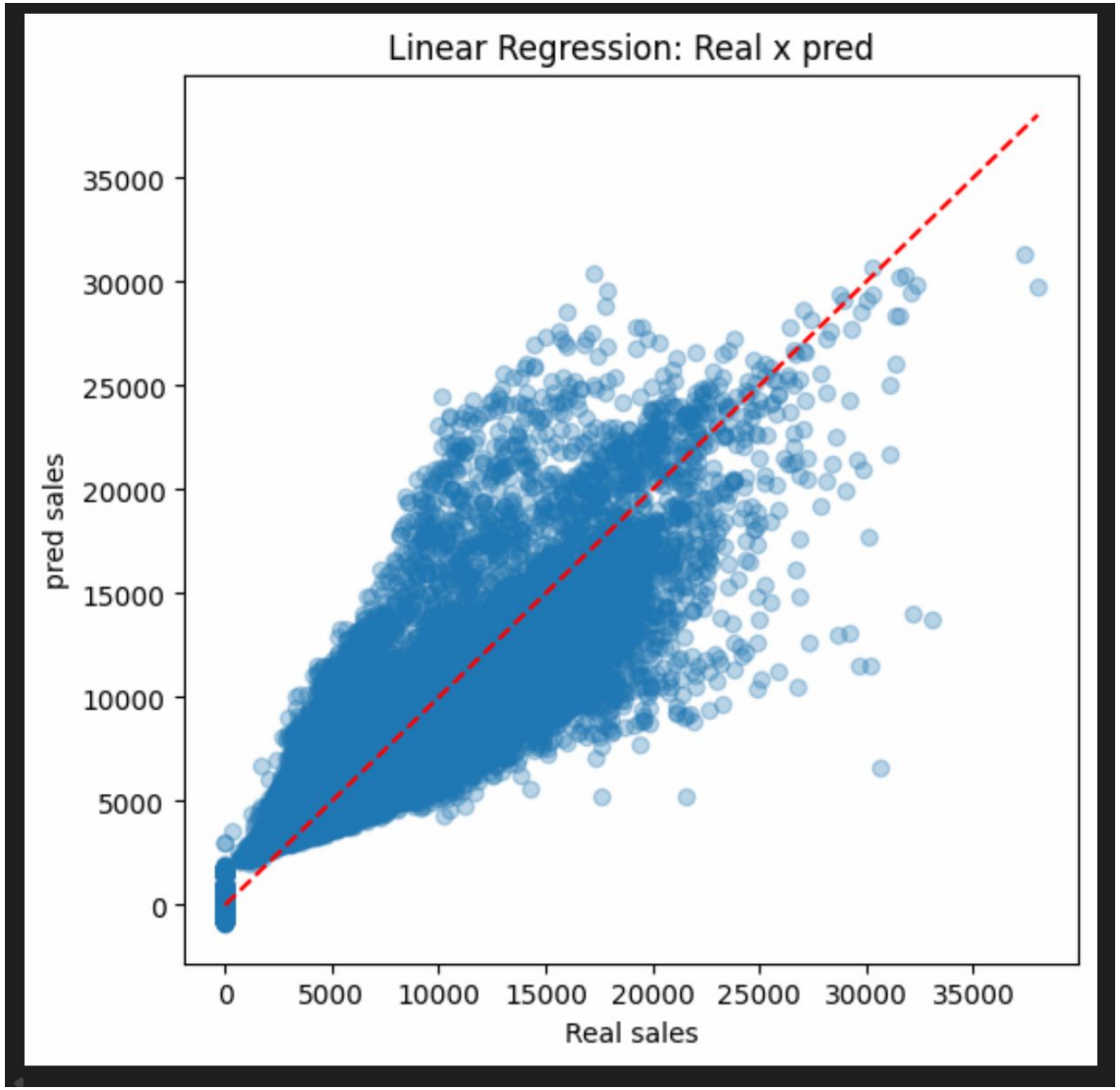
- Dropped the date column since it was an object(non numerical).
- Converted the “State Holiday” column to numerical data and dropped the original column
  - Checked for missing values.

Model

- We assumed it was a Linear Regression because we wanted to predict a single value.

```
Mean Squared Error (MSE): 2200939.1483
R-squared (R²): 0.8511
Mean Squared Error (MSE): 2200939.1483
Mean Absolute Error (MAE): 980.2598
```

Results


Linear Regression: Real x pred
pred sales
Real sales
0
5000
10000
15000
20000
25000
30000
35000