Sales Forecasting

Using Regressors in Machine Learning

Problem statement

Forecast the company's sales using less than a year of data and a very limited number of influencing factors

Objectives

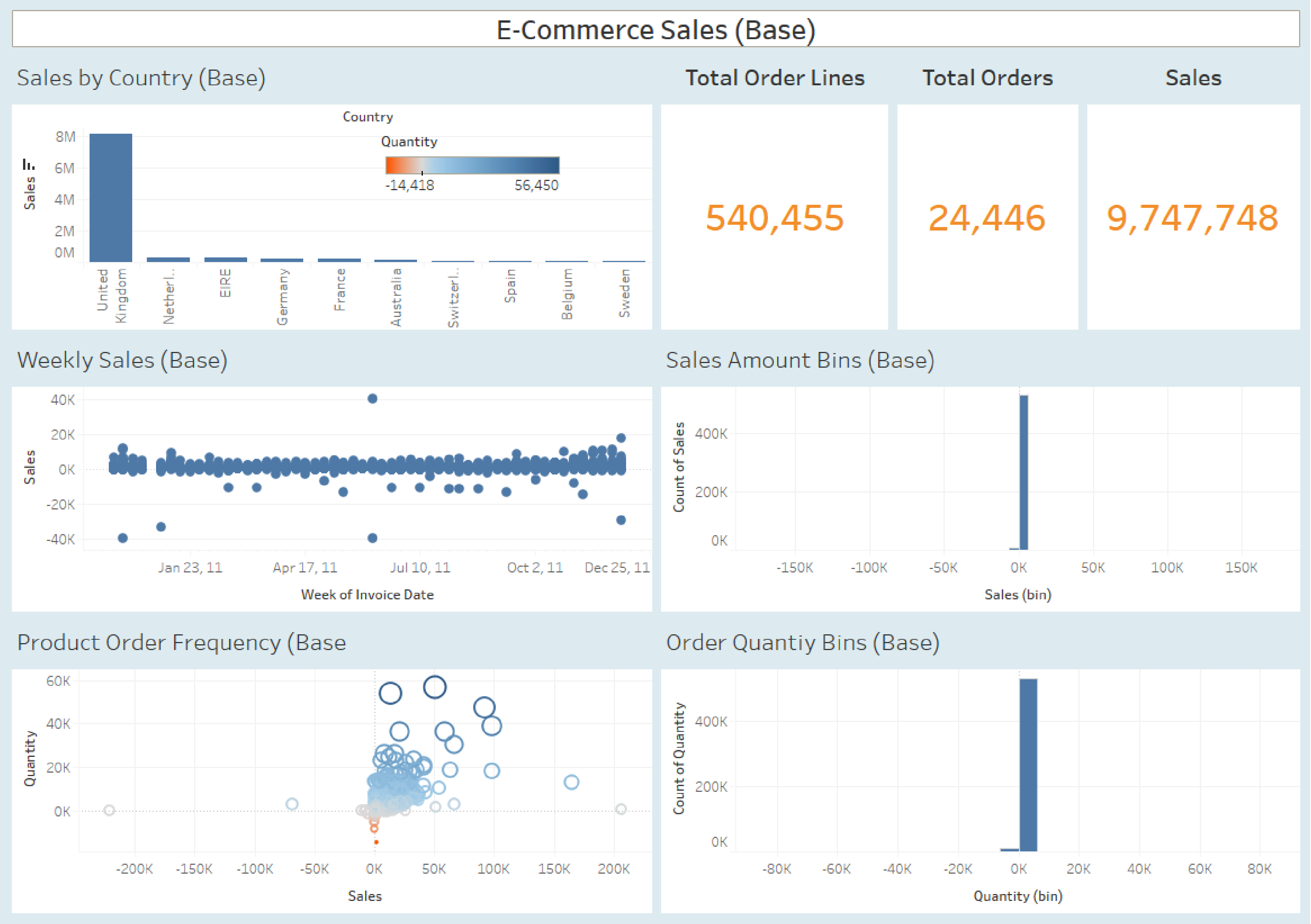
- Identify and ignore the noise in the data

- Build multiple regression models to find the best model and parameters to be used for future data

Exploratory Data Analysis

Base data

Details of data before clean-up



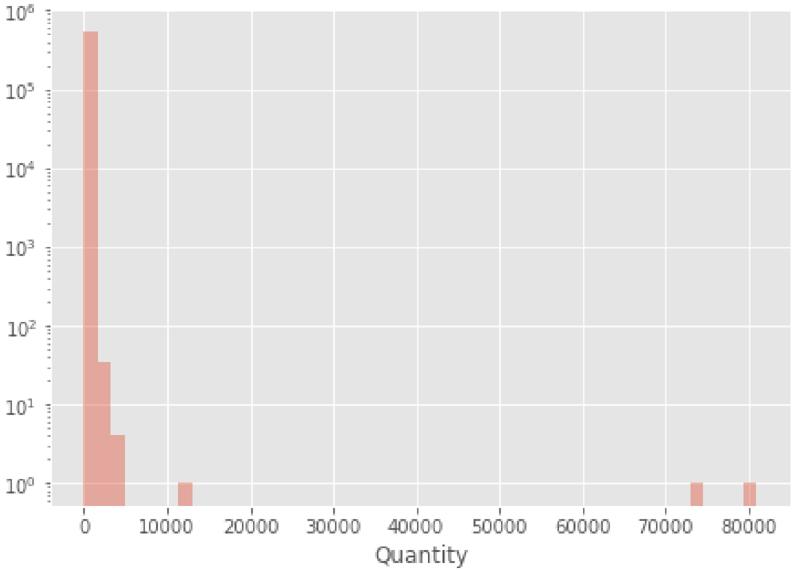
E-Commerce Sales (Base)

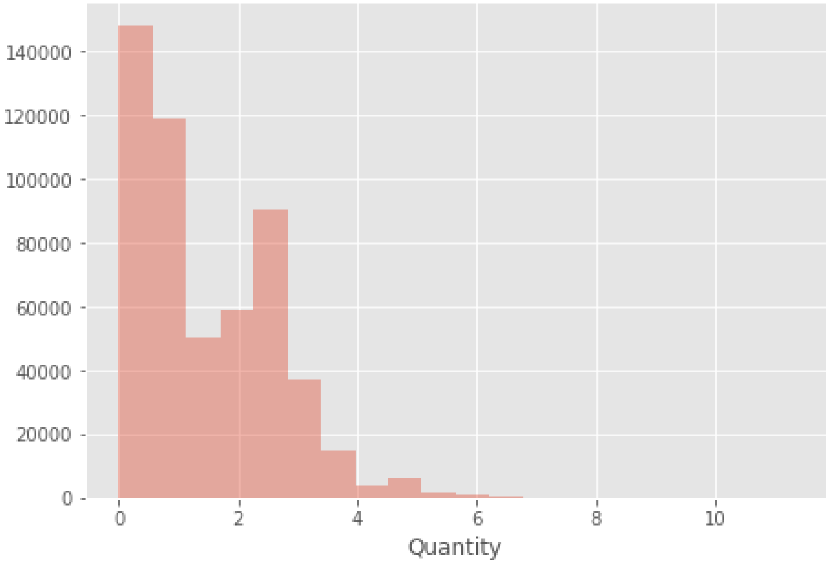
Total Order Lines: 540,455

Total Orders: 24,446

Sales: 9,747,748

Dealing with outliers





To determine where to cut the outliers:

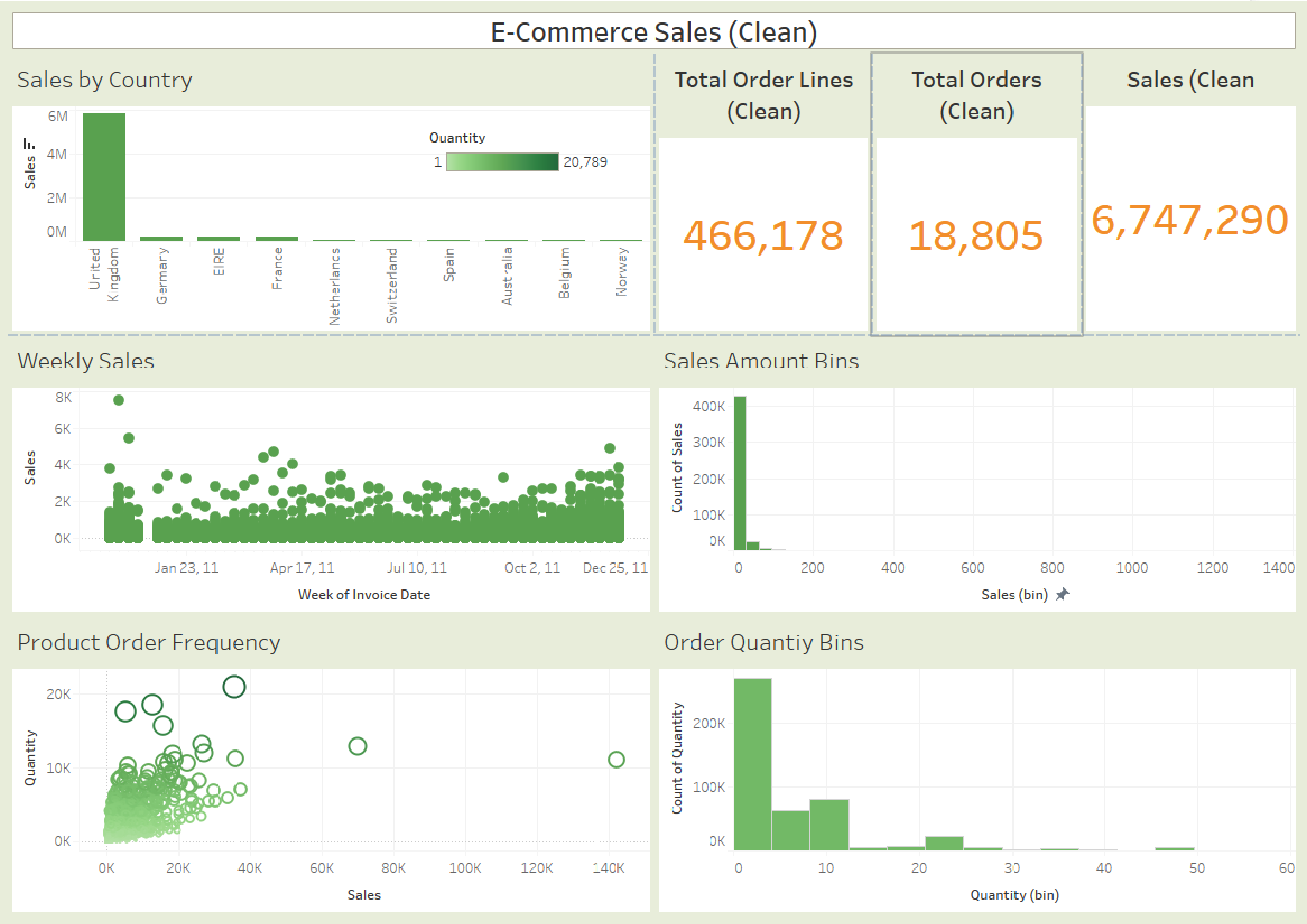
- 95% Quantile = 30 quantities

- exp(3) = 20.0855 ~ 20 quantities
- exp(4) = 54.5981 ~ 55 quantities
- exp(5) = 148.4131 ~ 148 quantities
- exp(6) = 403.4287 ~ 403 quantities

*The practical cutoff is ~ **55 quantities**

Clean data

Details of data after clean-up



E-Commerce Sales (Clean)

Total Order Lines (Clean)

466,178

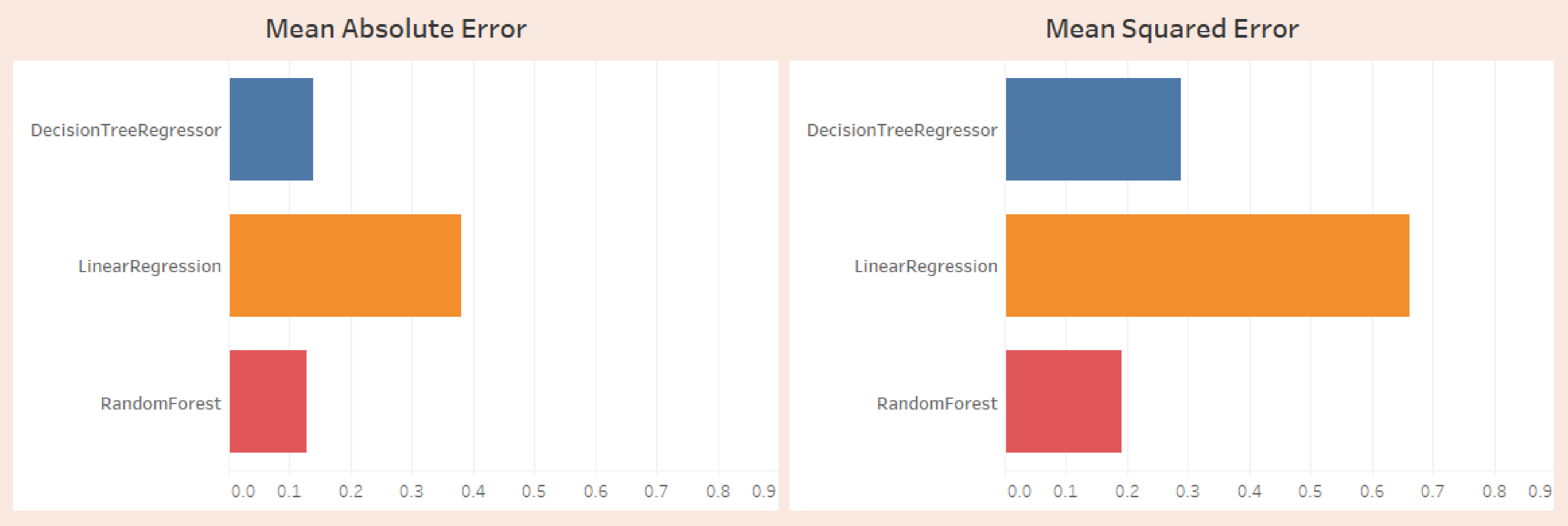
Total Orders (Clean)

18,805

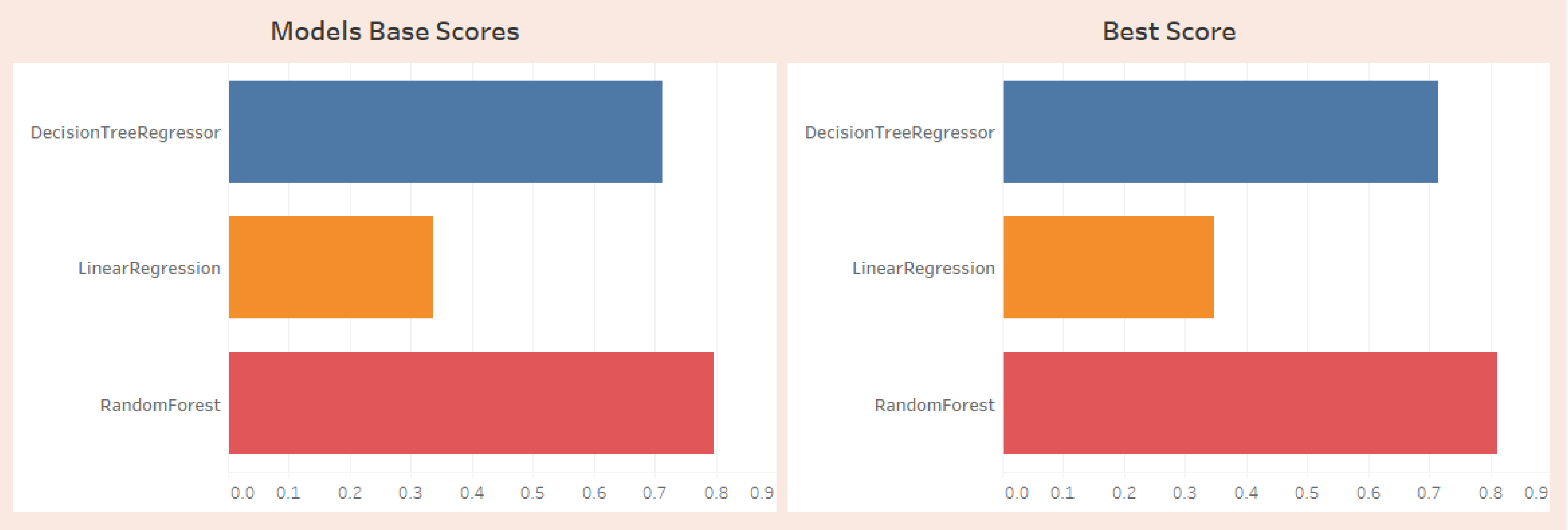
Sales (Clean

6,747,290

Quality of models



Model score



RandomForestRegression is the best model to use based on the base score and hypertuned score

Conclusion

With the help of python’s excellent data wrangling/cleaning libraries, the noise in the data can be easily taken cared of

Forecasting sales given the limited amount of information is still possible to achieve with high scores. These models need to be enhanced should new data with more features are available

Recommendation

With the limited number of features, a lot of things can still be done:

- Bucketize quantities and sales ranges
- Stock codes and descriptions can still be categorized to smaller groups
- Separate predictions between UK and the rest of the other countries
- Separate predictions on the returned/cancelled transactions
- Deep learning models to see if the best score can still be improved

Thank you