Hospital Case Study

Agenda

- Introduction
- Analytics
- Machine Learning
- Analysis

Introduction

Introduction

- This is a case study for Dubai Tourism that uses Hospital data
- The data has 248 records, and 24 columns.
- Therefore, there are 23 predictor variables, and one target variable, which is the total cost incurred by the hospital
- This is considered a very small dataset for machine learning purposes, and might not yield models with great accuracies (since data is very limited)
- The problem to solve here is to come up with a way to predict the total cost incurred by the hospital, given patient’s data

Analytics

Analytics

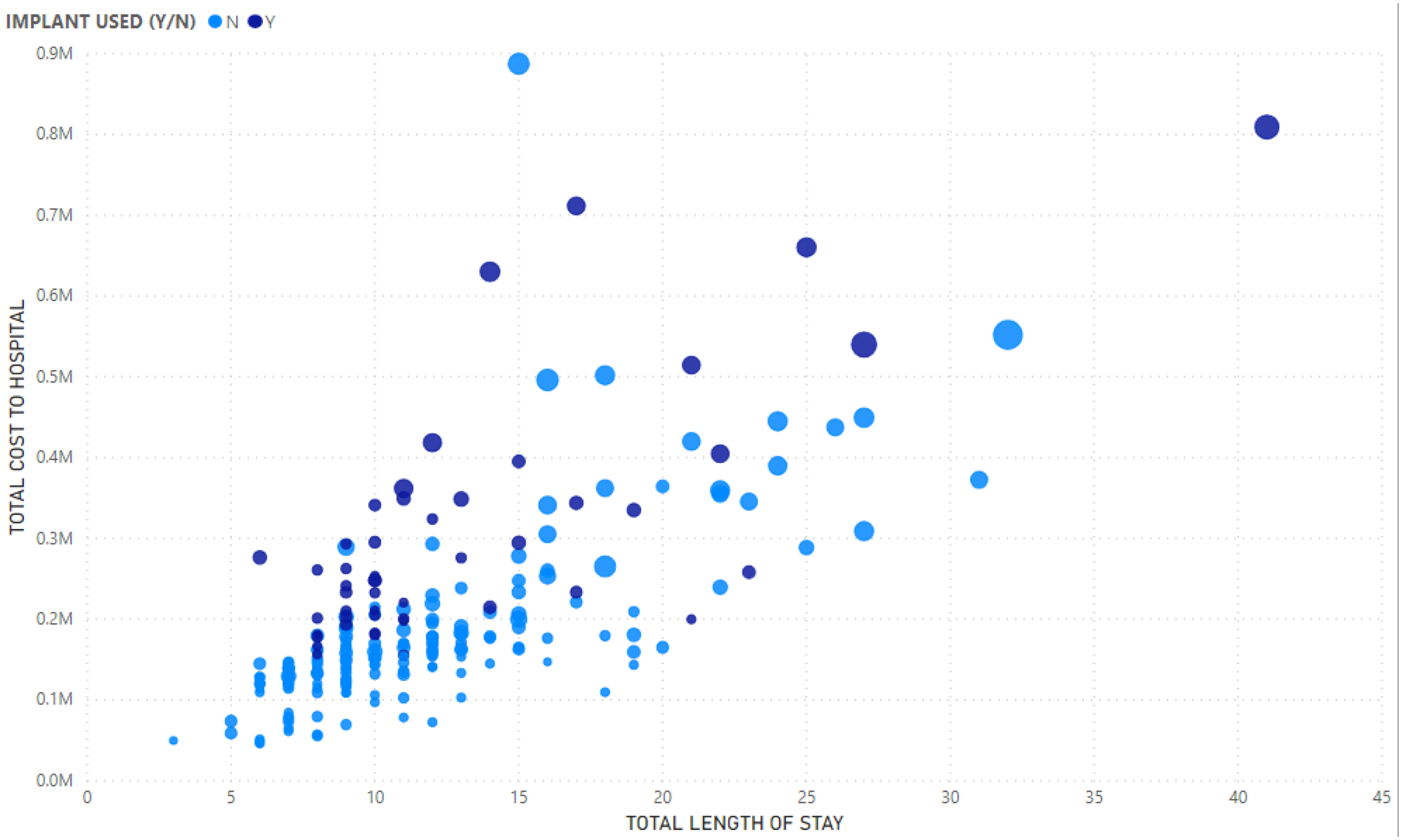
- In this section, we give an overview of the data in the form of analytics

- Multiple charts will be presented to give a deeper understanding of the data

- Plots have been generated using Microsoft PowerBI, and Python

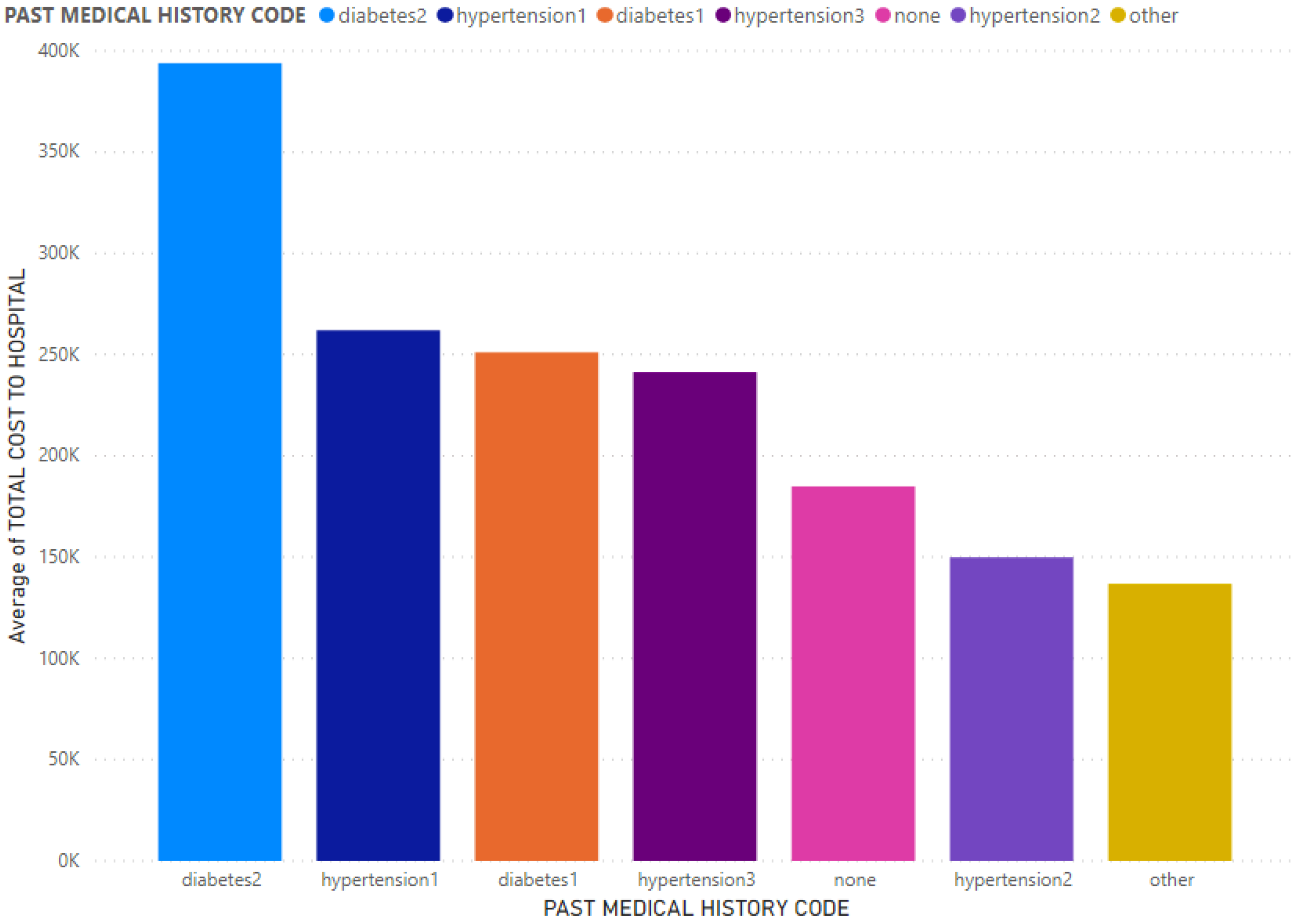
Total Length of Stay

- In this scatter plot, we show the total cost of hospital plotted agains the total length of stay in the hospital
- The dot size represents the total days spent in the ICU, and the dot color represents whether the patient has implants or not
- We can see from the graph that the bigger the dot (days in ICU), the higher the cost. And implants also generally indicate a higher cost to the hospital.



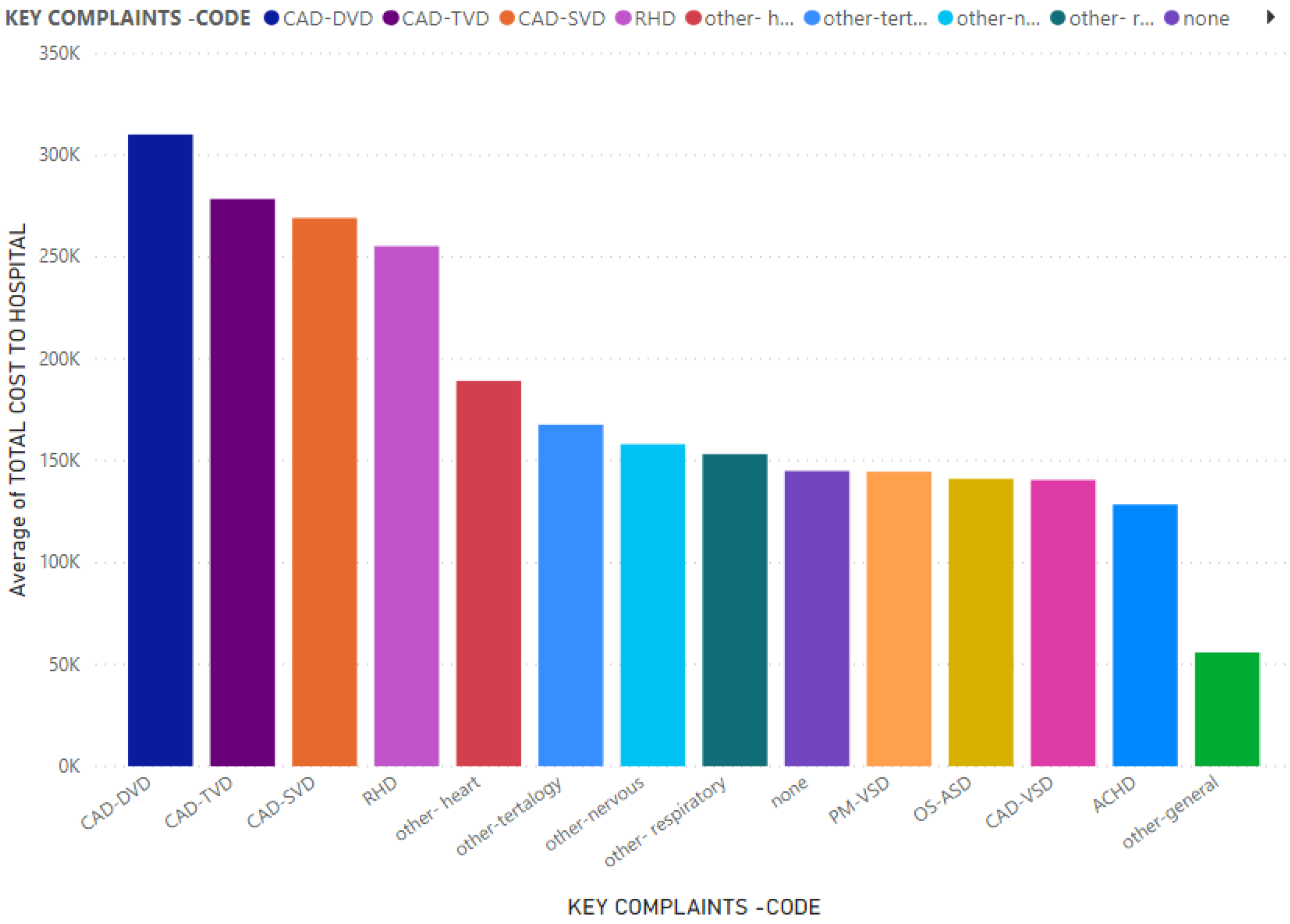
Past Medical History

- This graph gives us an idea about the different types of past medical history that patients have
- The impact of this history is very clear on the total cost to the hospital



Key Complaints

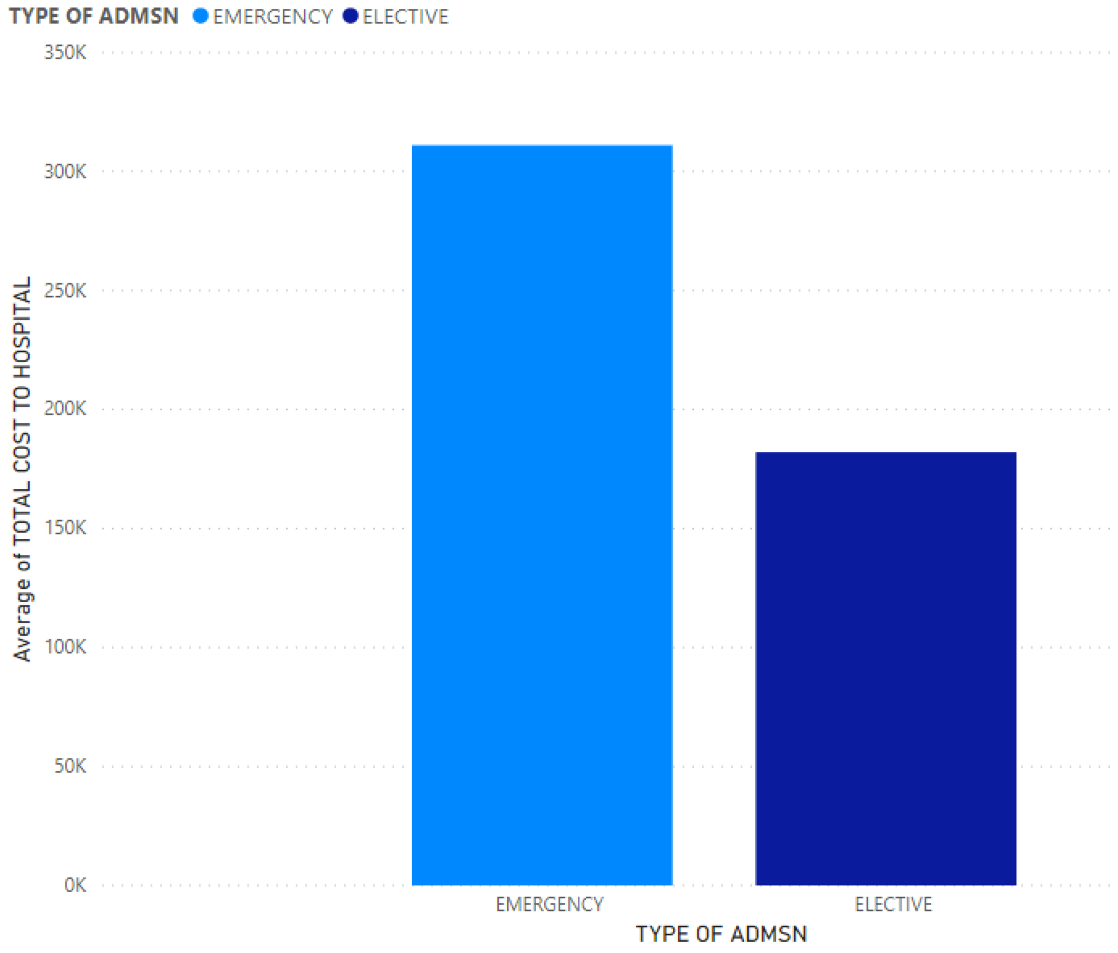
- This graph gives us an idea about the different types of complaints that patients have been admitted with
- The impact of the complaints to the total cost to the hospital is also very clear

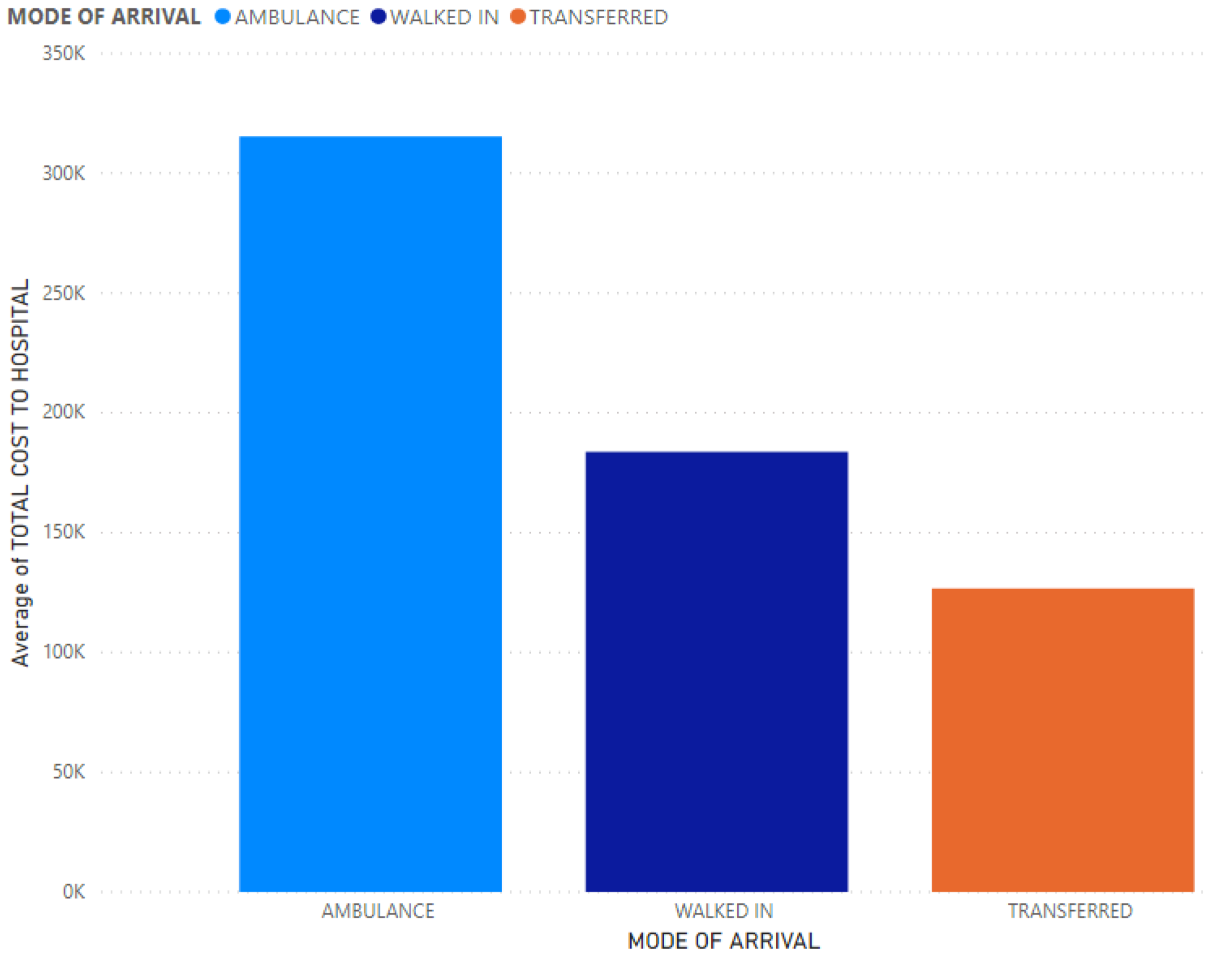


Admittance

- The way that patients have been admitted to the hospital (Emergency vs Elective), as well as the mode of transport used to get to the hospital has an impact on the cost to the hospital

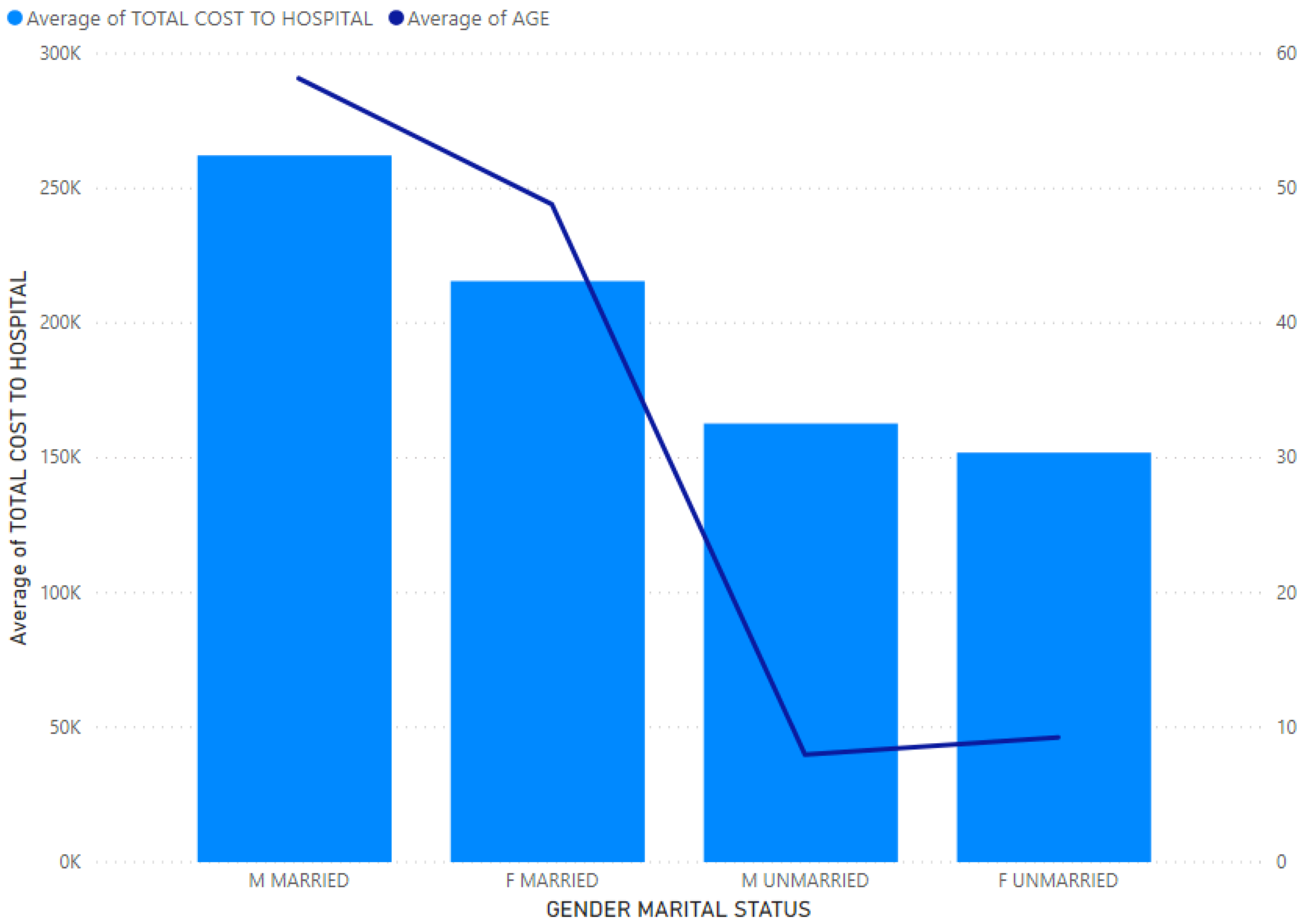
- As expected, using the ambulance is more expensive, and the more urgent the case, the more critical is the illness and therefore, we expect a higher cost





Gender, Age and Marital Status

- It seems that being married has a higher correlation with higher costs to the hospital, as well as being male
- The linear curve represents the average age in each category
- This is not good news for older married men, being the worst off demographic



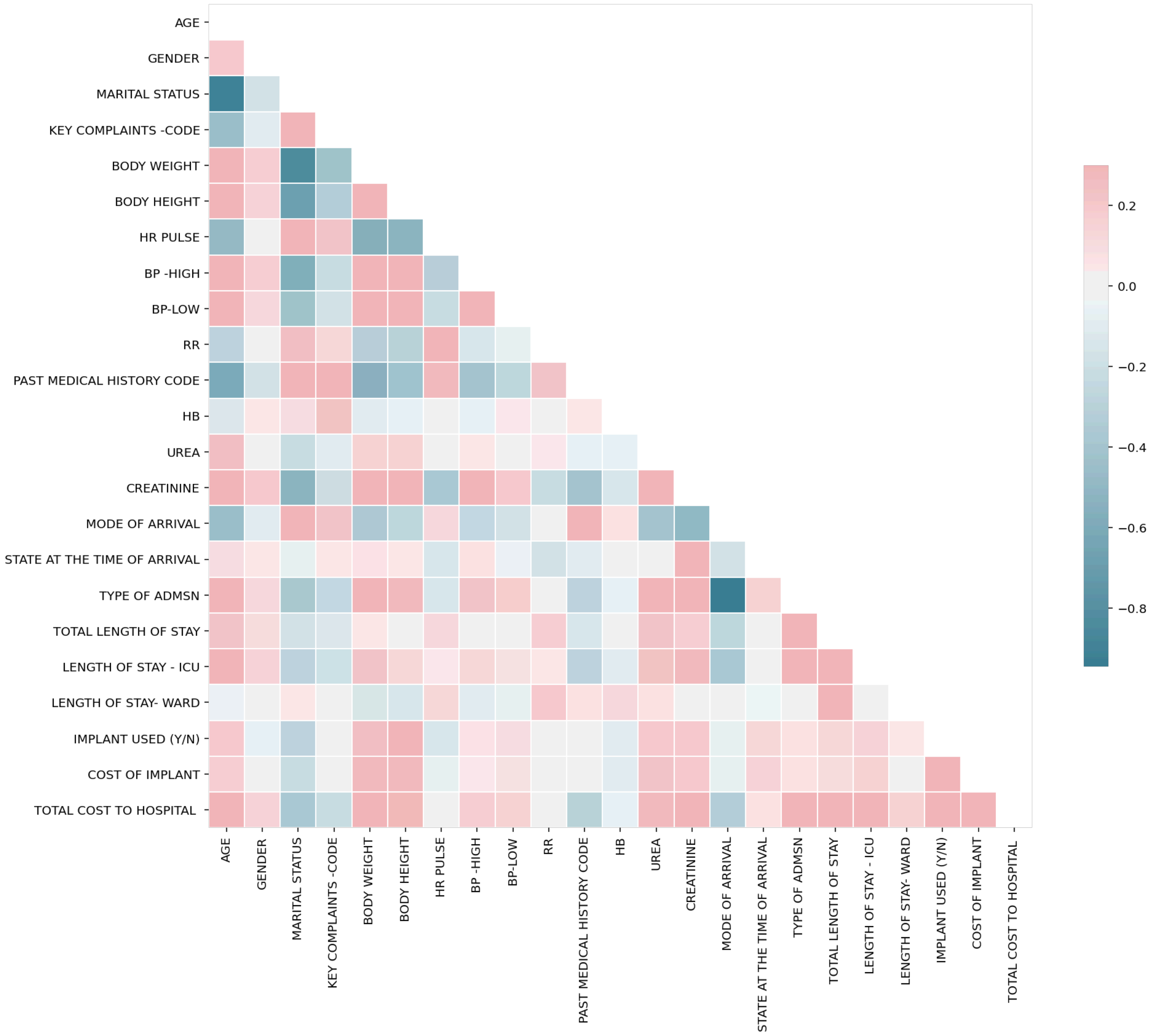
Correlation Matrix

- Using Python and Pandas, the following correlation matrix has been generated

- This visualization informs us of the correlations between the different columns of the data, whether this column is numerical or categorical (needs some processing to get the categorical columns correlations)

- We can see that there are fields that have higher correlations / impacts on the total cost to the hospital than others

- RR for example seems not to have any correlation with the cost, and therefore can be safely removed from the inputs to the regressions



Machine Learning

Machine Learning

- Python, Scikit-Learn, Pandas and a variety of libraries have been used for this case study
- Development was done using Jupyter Notebook
- The code is attached with this presentation (export Jupyter Notebook as HTML file)

Data Cleaning and Massaging

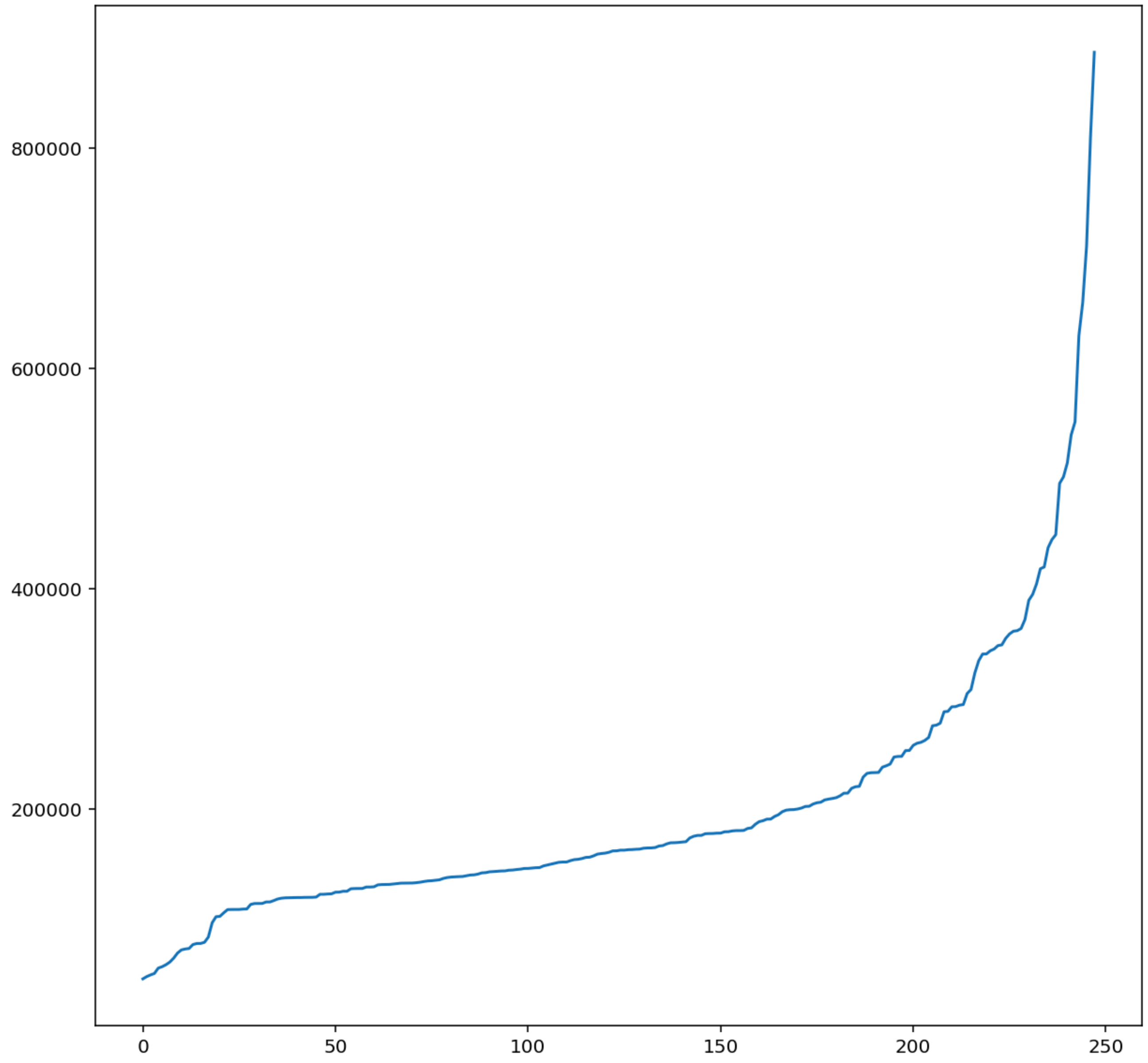
- Several data cleaning and massaging steps have been performed with Python

- Examples include imputing missing numbers, fixing one record with a problem age, finding problems with categorical data (such as lowercase / uppercase problems) and identifying outliers

- This has been well documented in the code in the Jupyter Notebook

Target Variable

- The first thing we do is to visualize the range of the target variable (total cost to hospital)
- We notice that the shape of the target variable resembles a "hockey stick", with a kink in the beginning (sub $100,000)
- This gives regressions a hard time in fitting the curve

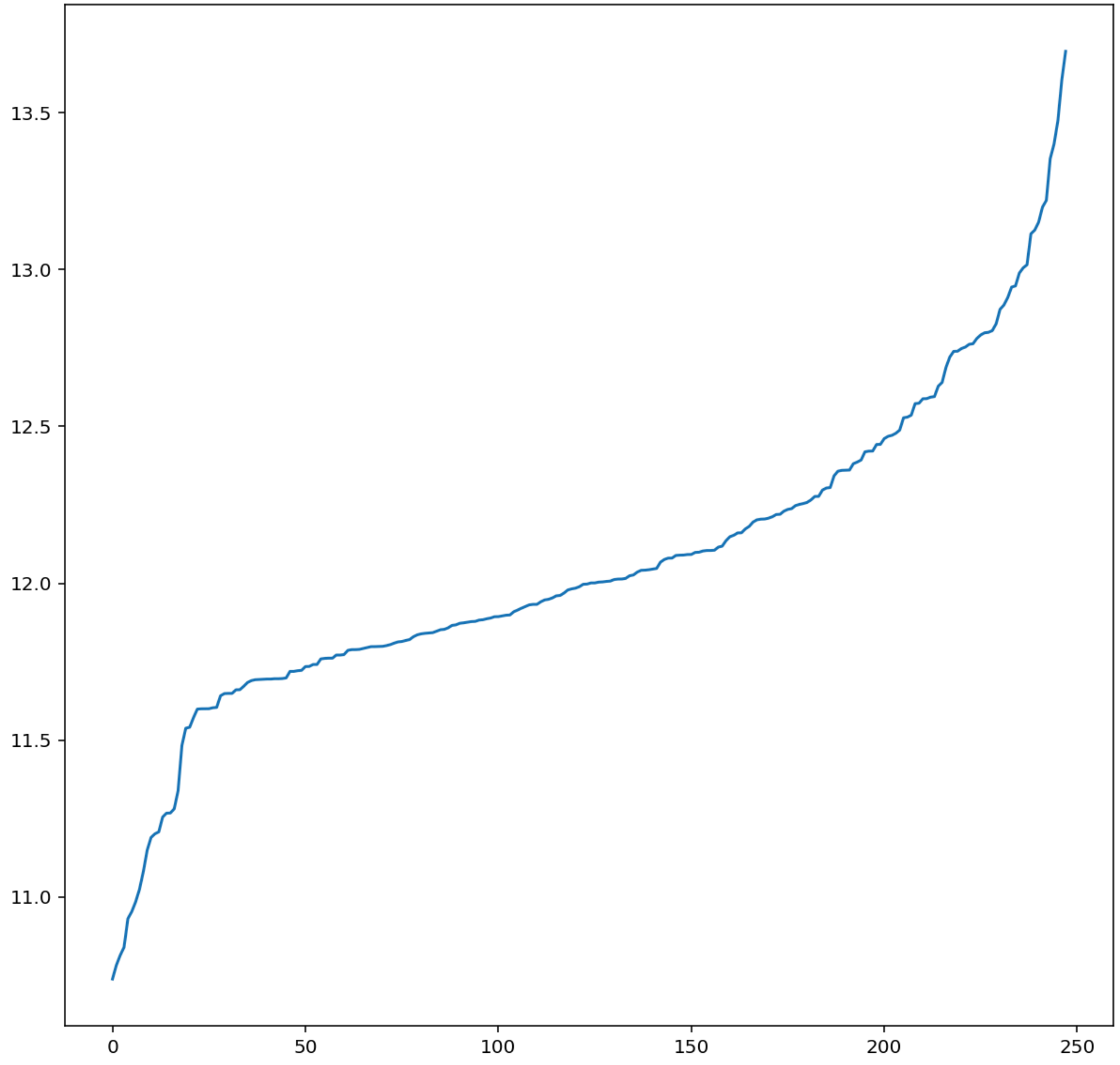


Log of Target Variable

- It would be better if we operated on the log of the target variable

- By taking the log, the curve is now smoother, and has a lesser range (between 11 and 13.6, rather than 50,000 and 880,000)

- It is now more linear, and regressions have a better chance fitting it



Preparing the Data

- What we do next is preparing the data for machine learning

- We split the data into train and test data (train 80% - test 20%), binarize all categorical columns (one-hot encoding), perform min-max on numerical columns, and add a square of all predictor numerical columns

- Since the target column still follows a "hockey stick" shape, even after the log, squaring predictor columns will provide better input, since now the regression can operate on polynomial-like inputs, e.g. a * x^2 + b * x + c

Building the Model

- After multiple trials, I settled on a StackingRegressor, with member regressors of Random Forest Regressor, Extra Trees Regressor, Linear Regression, and XGBoot Regressor

- The stacking of these regressors gives a really good performance, especially that the dataset is small and is not totally linear

- I'm tracking the R2 and Mean Squared Log Errors accuracy measures, to track performance

- I'm also using Grid Search with a 5-fold cross-validation splitting strategy, to look for the best combination of input parameters, and to choose the best model

Target Variable Outliers

- The target variable ranges from around 46,000 to 886,000
- However, inspecting and visualizing the data, there are only 18 records below 90,000, and 9 records above 500,000
- For the purpose of training the model, these could be considered as Outliers, and removed from the input training data, to improve the accuracy of prediction
- This totally depends on the business case. If these values CANNOT be considered as outliers by the business case, then we do another regression training at the end just to showcase the difference in accuracies (with and without those outlier records)

Analysis

Model #1 - Best Accuracy

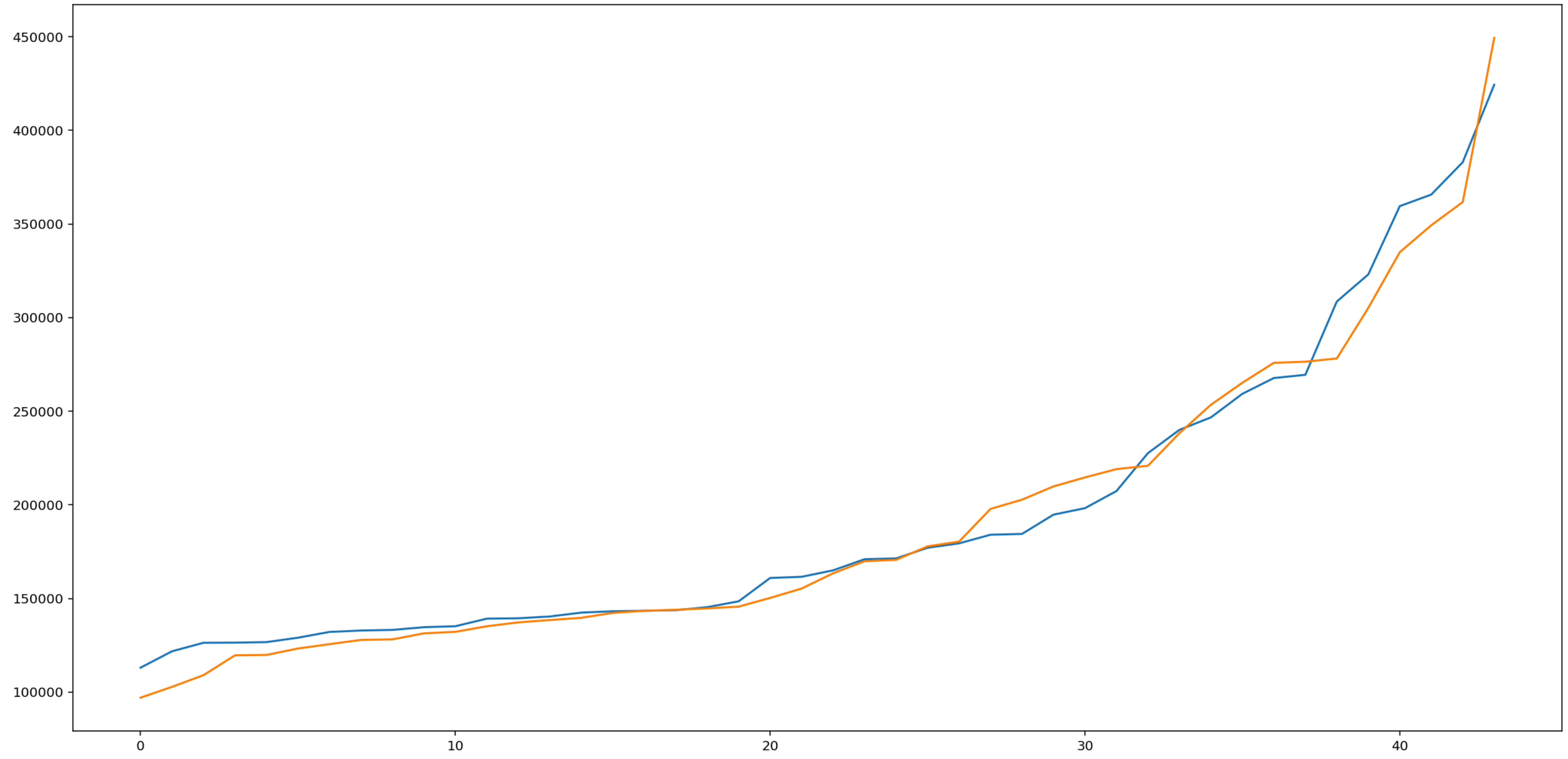
- Using the Grid Search and multiple ranges for target variable outlier filtering, the best accuracy has a R2 score of 91.3%

- Given the size of this dataset, this is great accuracy for a regression

- The target variable range was from 90,000 to 450,000. Everything outside this range is considered an outlier (27 records filtered out)

- The yellow line is the test dataset, and the blue line is the predicted values

- It can be observed how close the two curves are

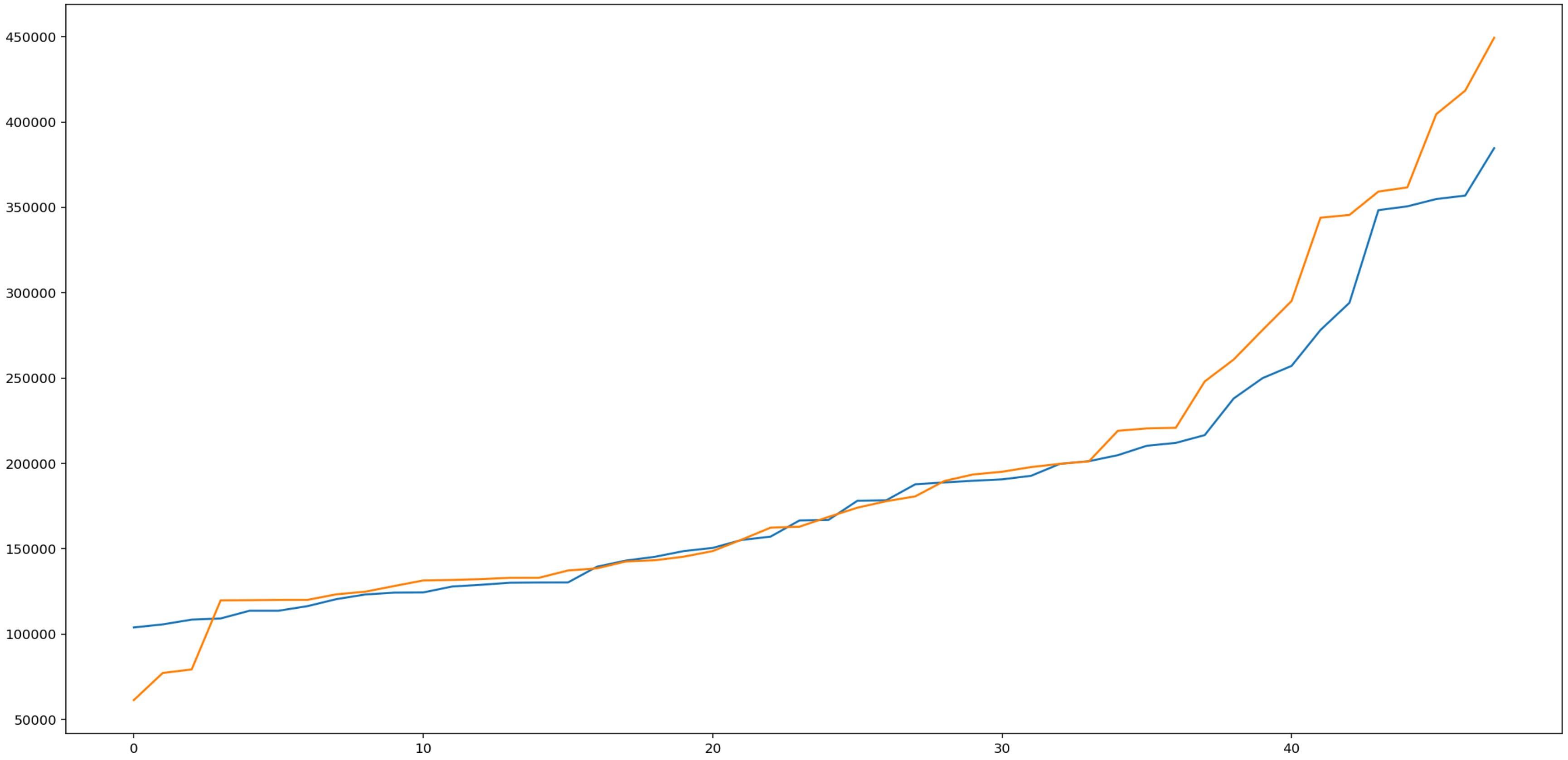


Model #2 - Next Best

- The next best model has a R2 score of 88.3%

- The target variable range was from 50,000 to 500,000. Everything outside this range is considered an outlier (only 12 records filtered out)

- The yellow line is the test dataset, and the blue line is the predicted values



Model #3 - Full Range for the Target Variable

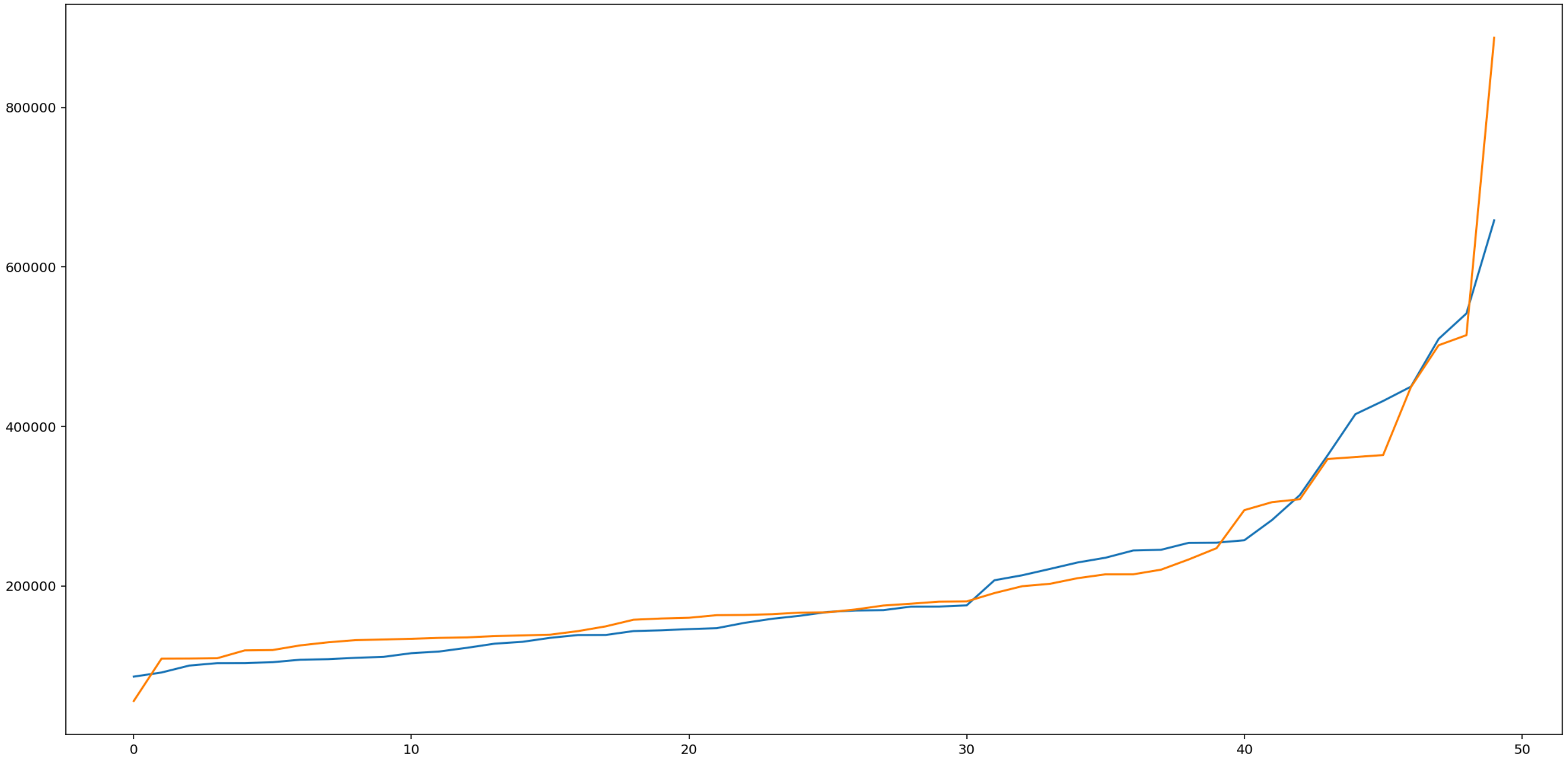
- Just to compare model accuracies, I ran the model using Grid Search on the full range of the target variable, using all records

- This model has an R2 score of 71.7%

- The model predicts up to around 650,000, when the real value is 880,000. This is not bad given that it needs to be determined whether these values are outliers or not

- The yellow line is the test dataset, and the blue line is the predicted values

- Even though the R2 score is not very high, but still the predictions are very close to the real values, given that it's not too close to the edges



Thank You