Thomas O’Gara

Part time data science at Flatiron

Mod 1 Project Blog post

**EDA – Exploratory Data Analysis:**

In the OSEMN data science process I thought the “EDA” phase was the most interesting for a number of reasons. First of all, it is the most “human” part of the data science process because it requires intuition and interpretation that may be harder for a machine or automated process to mimic. The first step in EDA is understanding and interpreting the distribution of the data. Some really helpful insights can be gained just by asking the right questions and making visualizations to show some general relationships.

The first step I took in my EDA was to look at some high-level descriptive statistics. I used Pandas built in .describe() and .info(). The .describe() method shows count, mean, median, min, max, and quartile values for every column in the DataFrame. This is useful for answering questions that involve general statistics such as variance, standard deviation, etc.

The next step I took was to visualize some distributions using .hist() method. This creates histograms showing the distribution for each column. Another way to view the distribution of a column is by building a Kernel Estimation Plot (KDE). This is usually overlaid on a histogram to create a line that visualizes the probability mass for every value in the histogram.

To check the linearity assumption, which is making sure each column has a linear relationship with the dependent variable, I used Joint plots. This is a scatterplot with the distributions of two different columns, a kde plot, and a simple regression line all on the same visualization. This was really useful for viewing the linearity assumption and distribution assumption for a predictor and target variable for a multiple linear regression model. I used the seaborn library to create the joint plots as follows:

*Sns.jointplot(x = <column>, y = <column>, data=<dataset>, kind=’reg’)*

The joint plots can also be useful for determining which variables should be numeric or categorical. If the scatterplot doesn’t follow a somewhat linear cloud pattern, and it looks more like columns on the chart it should probably be a categorical variable even if it isn’t completely intuitive at first.

If the distribution and linearity assumptions don’t hold very strong at first it doesn’t necessarily mean don’t proceed to model your data. A few of my columns didn’t follow a normal distribution or didn’t show a linear relationship with the target variable but I still proceeded to model my data and think about ways I can pre-process my data in following iterations. The EDA can also give you hints about how effective your data cleaning was. If there are severe outliers on your visualizations it may mean you want to go back and scrub your data again and try another iteration of EDA and modeling.

Another benefit of EDA is thinking about non-technical questions you may be asked in order to drive business decisions. In regards to my module 1 project and trying to predict home prices in King county I thought about a few questions that non-technical people may be interested in. One very obvious question is “which variables help predict the target variable?” You can show some simple visualizations about the relationship between a few key independent variables and the target variable to make your point. Another question I thought about was “what are the most expensive areas of King County?” This can be answered showing the home prices by zip code or by showing home prices by latitude or longitude on a map. These are some very simple insights that can be incredibly useful for business leaders without even going into your actual model.

One of the visualizations I used to show zip code and year built by home prices in King County was a boxplot using the following code:

*x = df['zipcode']*

*y = df['price']*

*plt.figure(figsize=(20,10))*

*sns.boxplot(x=x, y=y, data=df)*

*plt.title('Price by Zipcode')*

*plt.xlabel('Zipcode')*

*plt.xticks(rotation=90)*

*plt.ylabel('Price')*

*plt.show()*

Boxplots are a great visualization because they show you the median, which marks the mid-point of the data and it divides the box into two parts. The top of the box is the upper quartile (75%) and the bottom of the box is the lower quartile (25%). Then it shows you the upper whisker and lower whisker which show you the remaining upper and lower 25% of the data which is outside the middle 50% of the data.

EDA may also be very useful for narrowing your scope of predictor variables based on your initial visualizations. If a particular variable doesn’t show any kind of linear relationship with the target variable you may decide you want to drop the variable right then and there which can save you some time when iterating your linear regression model.

For me the big take away for Exploratory Data Analysis is choosing the right questions that a non-technical business person would care about and thinking about how to explain them showing visualizing and general statistics that anyone can understand. In the end the goal of data science to me is presenting your findings to non-technical stakeholders to drive business decisions.