Krystian Dennis

Follow 3 Followers About

House Price Predictions — EDA and Visualization in Python

Krystian Dennis Aug 14, 2019 · 4 min read ★

To be honest, visualizations are the main reason I chose to become a data scientist. I confess, I never read an article. I go straight to the infographics. They tell the whole story. If I can “see” the data, I can understand it.

For my first data science project for Flatiron School Online Data Science Boot camp, I analysed a data set for home sales in King County, Washington. The data set included information for sales between Sept. 2014-Sept. 2015. We were given information like square footage of living space, number of bedrooms, and even longitude and latitude for over 21,000 homes.

I followed the CRISP method for data mining and first took a look at my data set to make sense out of it. Once I had cleaned the data and completed feature engineering, I used Matplotlib and Seaborn to create visualizations of the dataset. I used pairplots, scatterplots, heatmaps, catplots and regplots to name a few. I am brand new to this, so I played around to find what worked best to accurately represent the data. I also played around colors and learned about html hex strings and matplotlib colormaps.

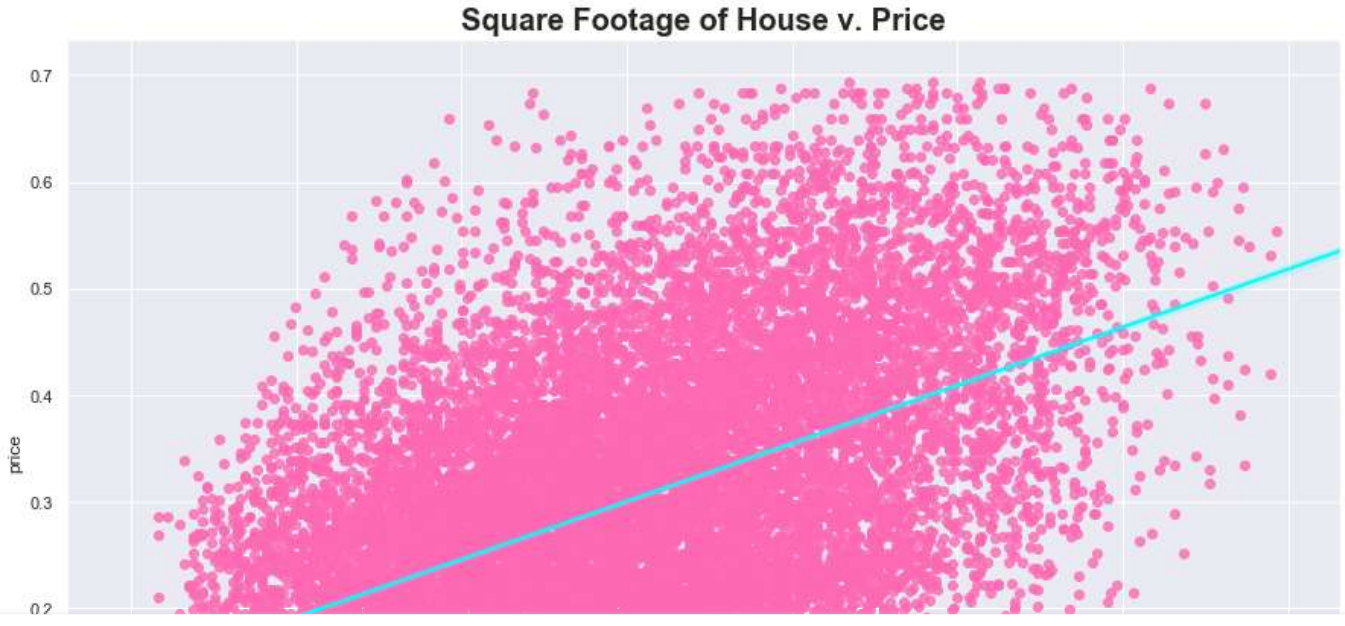
For my exploratory data analysis (EDA) I chose three variables from the dataset and examined how they affect the sales prices of homes.

1. **How does square footage affect sales price?**

In the dataset, there were several measures of square footage includes in the variables. I used the 'sqft_living' variable because it includes both above ground and basement square footage. I used a regplot for this visualization.

```
plt.figure(figsize=(15,10))
sns.regplot(x='sqft_living', y='price', data=df, color="#FF69B4", line_kws={"c
olor": '#00FFFF'})
plt.title("Square Footage of House v. Price", fontsize=20, fontweight='bold')
plt.show()

#regression line
```

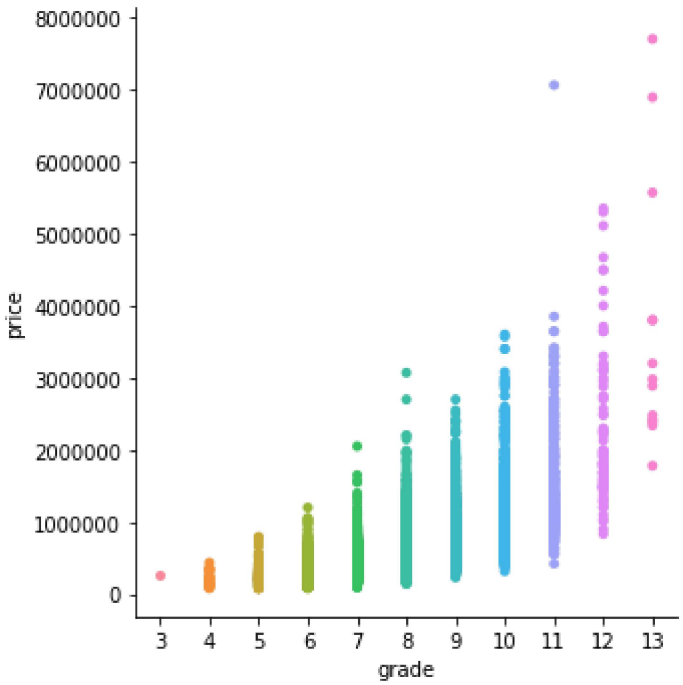


Not only did I find out that 'sqft_living' seems to have a strong linear relationship with 'price', but I used two html hex string colors. The data points are #FF69B4, also known as hot pink! My regression line is #00FFFF, or cyan.

2. **How does grade relate to price?**

The King County government created a grading system to rank all the home sales in the county. I used a catplot to see how this grading system affected the sales price of homes.

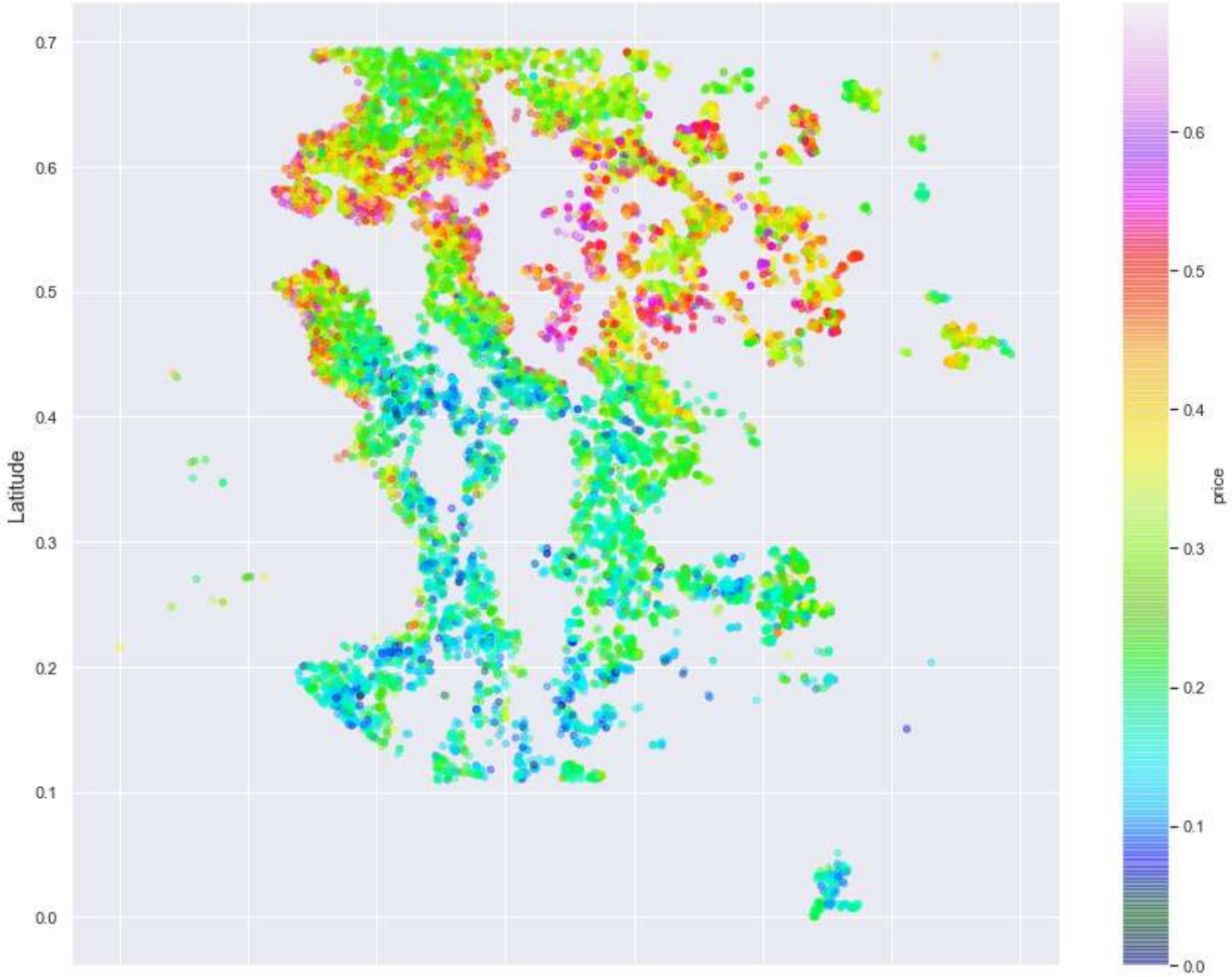
```
sns.catplot(x="grade", y="price", jitter = False, data=df);
```



Fortunately, Seaborn knows my heart. This catplot is rainbow colored all by itself. I can also see a clear linear relationship between 'grade' and 'price'. If a home received a grade of less than 6, it did not sell for over $100,000.

I used a scatterplot to visualize the connection between 'lat', 'long' and 'price'. This virtually allowed me to see how location in King County is reflected in the sales price of a home.

```
df.plot(kind="scatter", x="long", y="lat",c="price",
              cmap=plt.get_cmap("gist_ncar"),colorbar=True,
              alpha=0.4,figsize=(15, 12))
#labels
plt.ylabel("Latitude", fontsize=14)
plt.xlabel("Longitude", fontsize=14)
plt.show()
```



the plot. It was later confirmed by correlation and regression that 'lat' has a stronger impact than 'long' on sales price. I also changed the matplotlib colormap to "gist_ncar". The vivid color array allows for the difference in price point to be clearly seen.

Using Matplotlib and Seaborn to create visualizations of the dataset helped me get an understanding of how the predictor variables impact price. The visualizations helped me to answer important questions about the dataset and determine which variables to include for analysis.

Data Science