terenceshin.medium.com

Exploratory Data Analysis — Visualizing Variables (Part 2/2)

Terence Shin

6–8 minutes

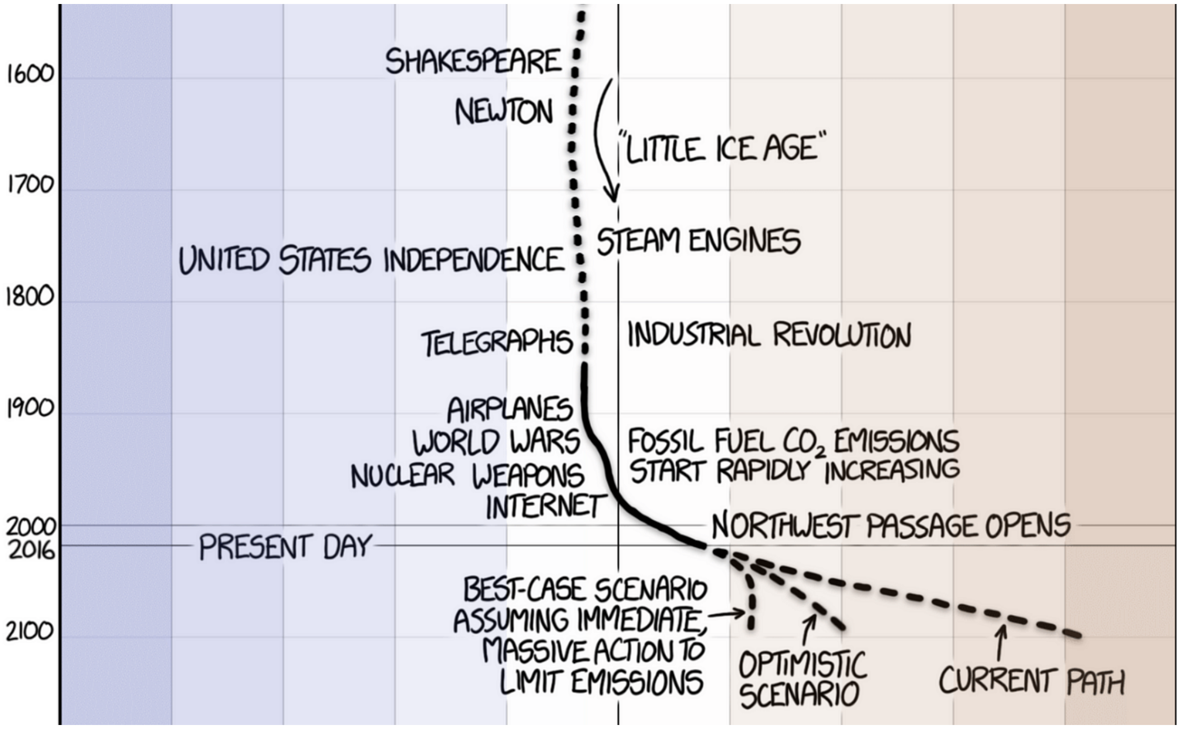
Week 4 of 52

A picture equals a thousand words…

And that's why everyone loves data visualizations, even those that say they hate numbers. Why? Because if done properly, it's an effective way to quickly communicate a lot of information in a short period of time — and time is everything for us Millenials and Gen Z's ;).

To give an example, the image below is only a snippet of the entire visualization, but this graph visualizes how fast we've increased the Earth's average global temperature in such a short period of time (see the full graph here). I think this is more impactful and insightful than stating that "the Earth's temperature is the highest it's ever been and is increasing the fastest that it ever has historically." Or maybe it isn't — but the graph certainly compliments the statement above if anything.

1500
CE
PRINTING PRESS
COLUMBUS
EUROPEAN RENAISSANCE



Visualization of Earth’s temperature over time

Recap from my last post

Quickly reviewing what I said in my last post, there are three main components of exploratory data analysis:

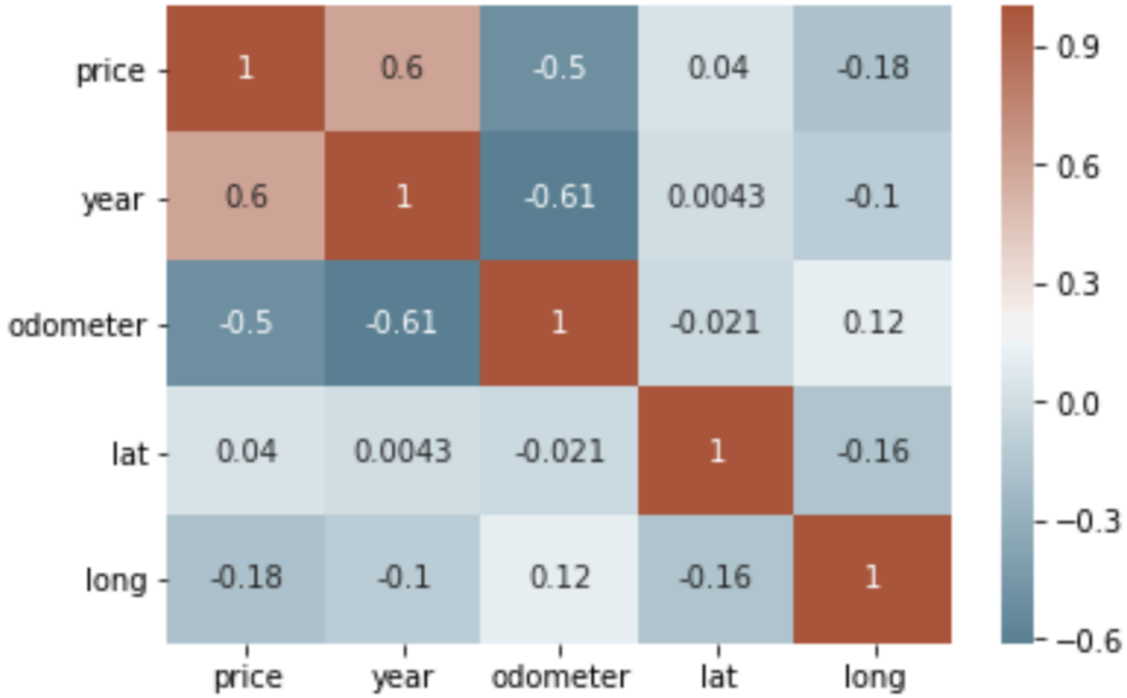
1. Understanding your variables
2. Cleaning your dataset
3. **Analyzing relationships between variables**

Previously, we went over the first two points. In this post, we’ll revisit the Used Car dataset and focus on the third component of EDA. Let’s dive into it!

Correlation Matrix

The first thing I like to do when analyzing my variables is visualizing it through a correlation matrix because it’s the fastest way to develop a general understanding of **all** of my variables. To review, **correlation** is a measurement that describes the relationship between two variables — if you want to learn more about it, you can check out my statistics cheat sheet here.) Thus, a **correlation matrix** is a table that shows the correlation coefficients between many variables. I used **sns.heatmap()** to plot a correlation matrix of all of the variables in the used car dataset.

```
# calculate correlation matrix
corr = df_cleaned.corr()# plot the heatmap
sns.heatmap(corr, xticklabels=corr.columns,
yticklabels=corr.columns, annot=True,
cmap=sns.diverging_palette(220, 20, as_cmap=True))
```



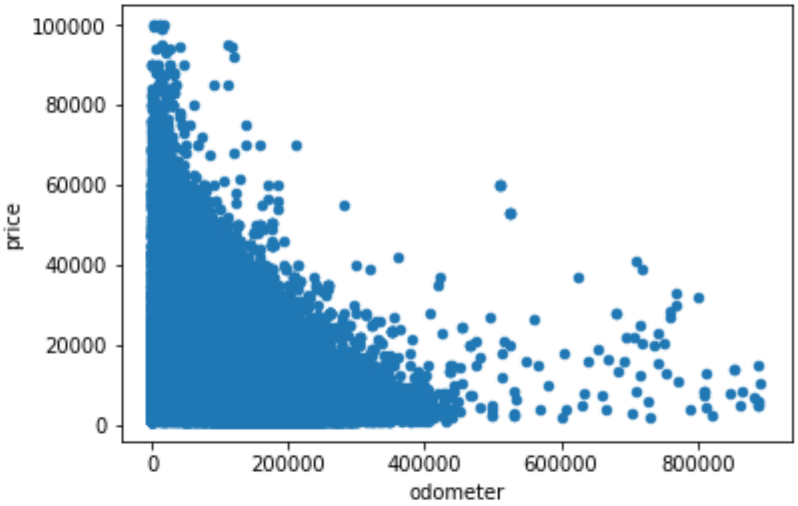
We can see that there is a positive correlation between price and year and a negative correlation between price and odometer. This makes sense as newer cars are generally more expensive, and cars with more mileage are relatively cheaper. We can also see that there is a negative correlation between year and odometer — the newer a car the less number of miles on the car.

Scatterplot

It’s pretty hard to beat correlation heatmaps when it comes to data visualizations, but scatterplots are arguably one of the most useful visualizations when it comes to data.

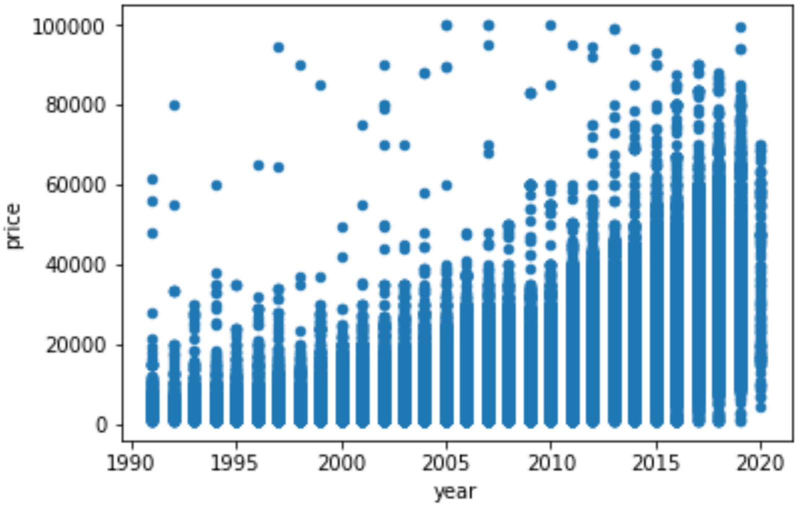
A scatterplot is a type of graph which ‘plots’ the values of two variables along two axes, like age and height. Scatterplots are useful for many reasons: like correlation matrices, it allows you to quickly understand a relationship between two variables, it’s useful for identifying outliers, and it’s instrumental when polynomial multiple regression models (which we’ll get to in the next article). I used **.plot()** and set the ‘kind’ of graph as **scatter.** I also set the x-axis to ‘odometer’ and y-axis as ‘price’, since we want to see how different levels of mileage affects price.

```
df_cleaned.plot(kind='scatter', x='odometer', y='price')
```



This narrates the same story as a correlation matrix — there’s a negative correlation between odometer and price. What’s neat about scatterplots is that it communicates more information than just that. Another insight that you can assume is that mileage has a diminishing effect on price. In other words, the amount of mileage that a car accumulates early in its life impacts price much more than later on when a car is older. You can see this as the plots show a steep drop at first, but becomes less steep as more mileage is added. This is why people say that it’s not a good investment to buy a brand new car!

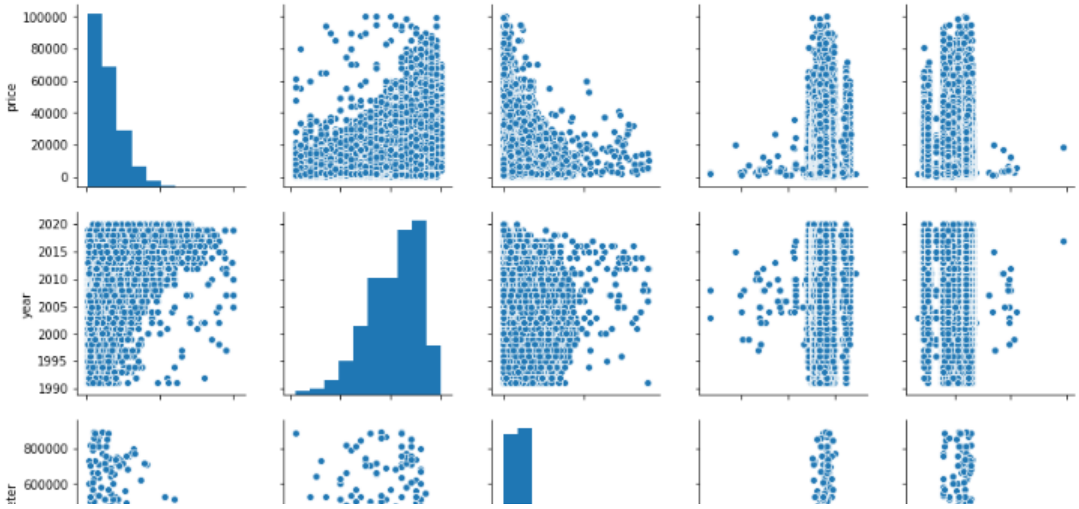
```
df_cleaned.plot(kind='scatter', x='year', y='price')
```

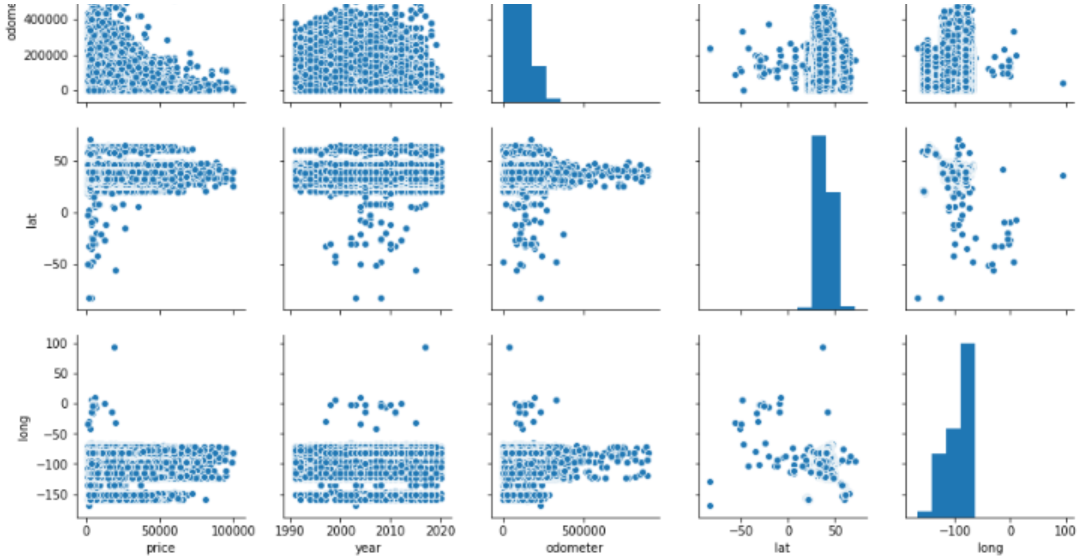


To give another example, the scatterplot above shows the relationship between year and price — the newer the car is, the more expensive it’s likely to be.

As a bonus, **sns.pairplot()** is a great way to create scatterplots between all of your variables.

```
sns.pairplot(df_cleaned)
```

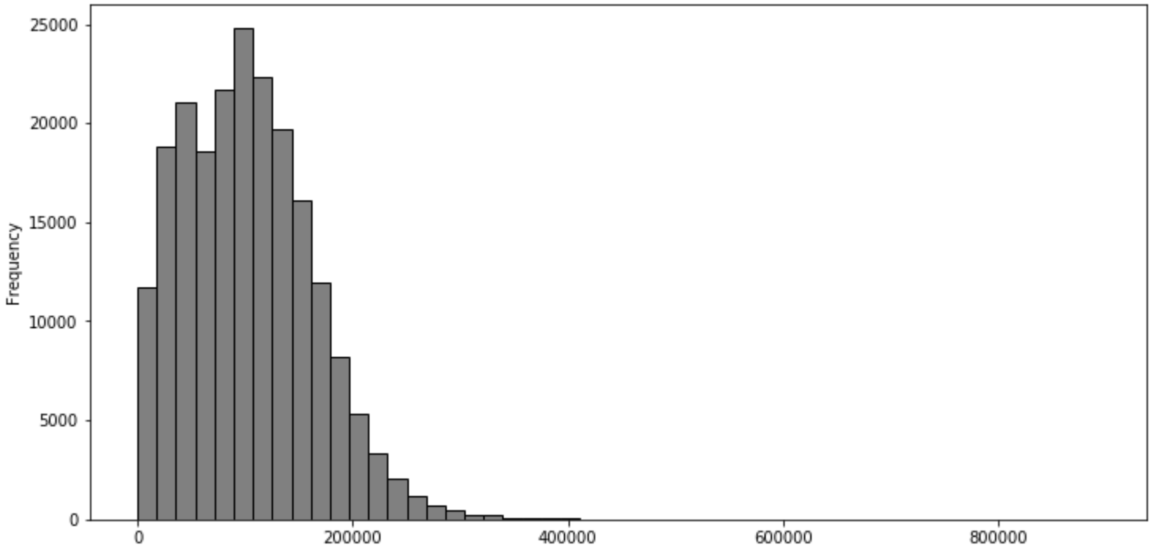




Histogram

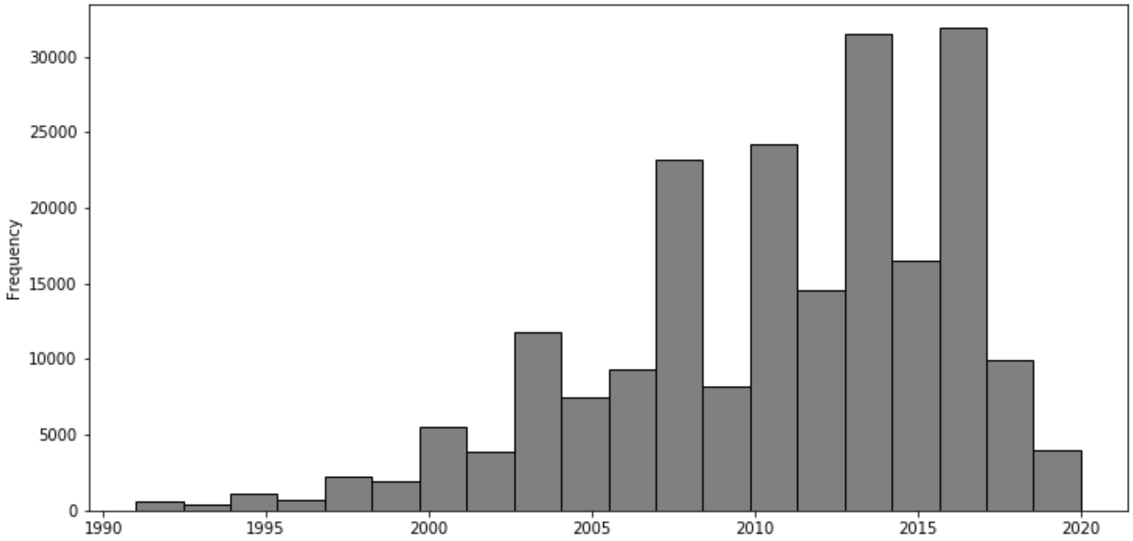
Correlation matrices and scatterplots are useful for exploring the relationship between two variables. But what if you only wanted to explore a single variable by itself? This is when histograms come into play. Histograms look like bar graphs but they show the distribution of a variable's set of values.

df_cleaned['odometer'].plot(kind='hist', bins=50, figsize=(12,6), facecolor='grey',edgecolor='black')df_cleaned['year'].plot(kind='hist', bins=20, figsize=(12,6), facecolor='grey',edgecolor='black')



df_cleaned['year'].plot(kind='hist', bins=20, figsize=(12,6),

```
facecolor='grey',edgecolor='black')
```

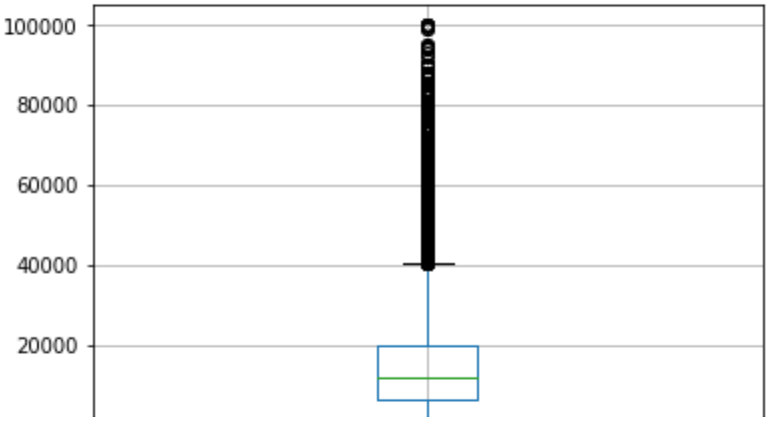


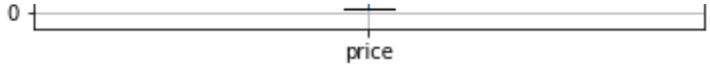
We can quickly notice that the average car has an odometer from 0 to just over 200,000 km and a year of around 2000 to 2020. The difference between the two graphs is that the distribution of ‘odometer’ is positively skewed while the distribution of ‘year’ is negatively skewed. Skewness is important, especially in areas like finance, because a lot of models assume that all variables are normally distributed, which typically isn’t the case.

Boxplot

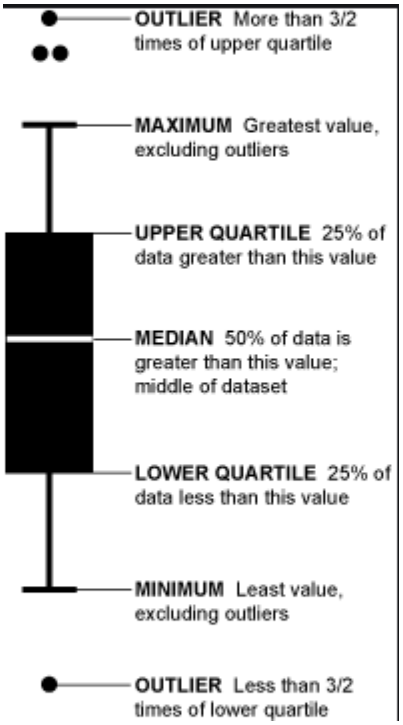
Another way to visualize the distribution of a variable is a boxplot. We’re going to look at ‘price’ this time as an example.

```
df_cleaned.boxplot('price')
```





Boxplots are not as intuitive as the other graphs shown above, but it communicates a lot of information in its own way. The image below explains how to read a boxplot. Immediately, you can see that there are a number of outliers for price in the upper range and that most of the prices fall between 0 and $40,000.


OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

There's several other types of visualizations that weren't covered that you can use depending on the dataset like stacked bar graphs, area plots, violin plots, and even geospatial visuals.

By going through the three steps of exploratory data analysis, you'll have a much better understanding of your data, which will make it easier to choose your model, your attributes, and refine it overall.

You can see the full code here.