This project demonstrates how to use python for data analytical purposes. Here, the sales data of used cars around the country is extracted and evaluated, so that any average trends can be seen. Then, the sales data for Porsche specific cars goes through the same process, and then compared to the first , so that an analysis can be made on how to improve upon certain figures compared to the regular market.

The notebook starts by mounting Google Drive to access the files on it. The data file containing the used car sales data, which is found on Kaggle, was already uploaded to google drive.

- Uploading from Google Drive is much more efficient than waiting 20 minutes for a large csv file to upload from the hard drive, due to the file being so large.

The data is loaded from a file in CSV format to a pandas DataFrame named 'df'.

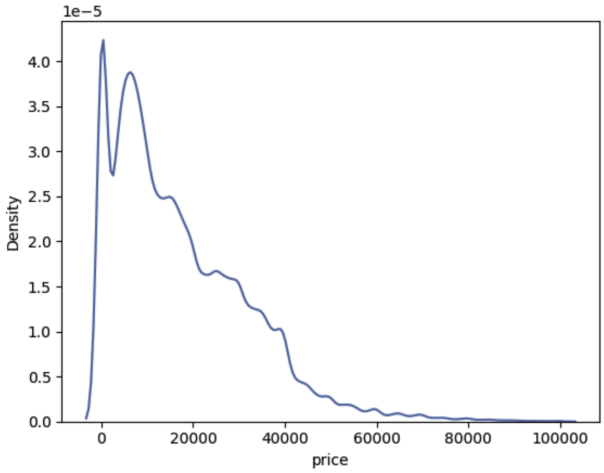
Basic data inspection is done by viewing the DataFrame by using 'df'.

Different plotting libraries, such as Matplotlib and Seaborn, are imported for the visualization of data. Some preliminary plots are boxplots and KDE (Kernel Density Estimation) plots to have an understanding of the price distribution.

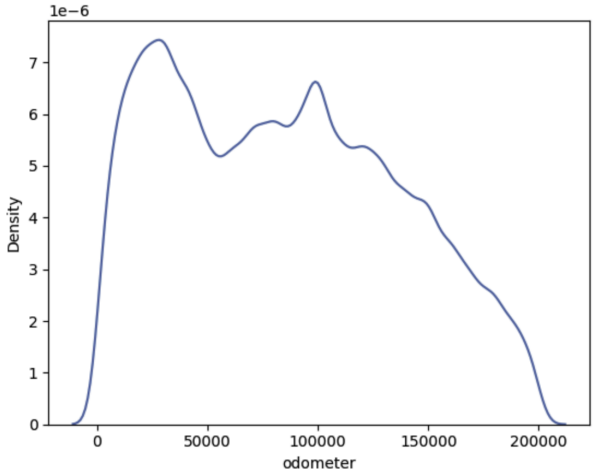
Outlier removal is a big focus. The first few plots are clearly showing skewed data, mainly caused by incorrect entries in the data set. Prices above certain thresholds are filtered out in several stages, as is similar filtering on 'odometer' readings, to focus on more typical values. After filtering outliers, more correct and readable plots can be derived from the data.

- Many of the incorrect data points are price/odometer readings which are 0, 1, 100, 456789, 999999, etc.

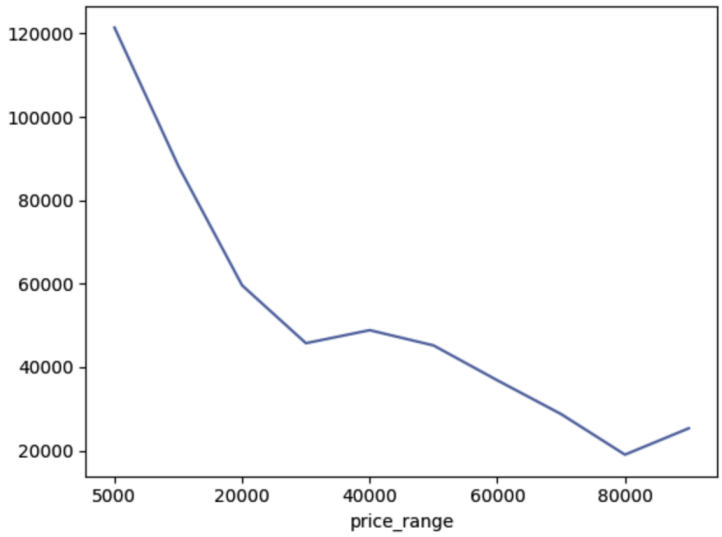
Once the data sets are cleaned enough to derive data from, different data points can be plotted to visualize the general trends for each category. In the chart below, the density of cars for sale at different price points are plotted. It can be seen that the majority of cars are sold in the $10,000-15,000 range. This can show dealers the price range to target on cars so that they would be easier to buy and sell, and eventually be profitable for their business.



The same process is done with the odometer readings. Once the outliers are removed, the density of cars for sale with different mileage are plotted. It can be seen that cars with either 25,000 or 100,000 are sold a lot more than cars at other mileages. This can also help dealerships determine whether a car would sell for a desired price. Many times mileage is a deciding factor on whether or not someone will buy a vehicle, and this analysis shows that a car with 25,000 or 100,000 miles should be easier to sell than a car with different mileage.



After this, there needs to be a way of comparing the main values, price and odometer. If every single data point of price x mileage was plotted, the plot would be a mess of points, messy, unreadable, and quite useless. So in order to properly plot these values, first the price points must be grouped into ranges. After each price range is set, all the odometer values within the range must be added so that the value can be averaged. After the odometer readings are averaged, there is only one point per price range, and the plot can be created in a neat and understandable manner. As expected, it can be seen from the newly created graph that the higher mileage the car, the lower of a price it will sell for. The graph continues its linear decline as expected up until about $40,000, showing that as a car drives more and increases its mileage, the sale price continues to decline. However, it can be seen that there is a slight spike around $40,000- 50,000. This shows that cars priced around that point, still sell even though the mileage may be higher. An analysis that can be made from that is that the $40-50k price point is around the price of many used luxury cars, and many overlook the fact a car might have higher mileage just so that the “luxury” car can be purchased. Meaning, a used luxury car with higher mileage that is priced on the higher end will be more likely to sell than that of a “normal” car.



The data is regrouped based on the region to determine the average price together with other regional statistics.

Distributions of filtered price and filtered odometer value are represented with the help of histograms and KDE plots.

Another data set that focuses exclusively on used Porsche cars is uploaded.

Same data cleaning and visualization steps are performed on this dataset of Porsches.

The Porsche data set is compared with the general data set to visualize the differences between the data sets, and evaluate any visible trends or trend differences.

Looking at figure 15, the plot shows the average odometer reading for different price ranges in the car market. This is extremely useful data, as it shows the price trends for each odometer range, which is very important when deciding the market value of the car.

Looking at figure 18, you can see the average odometer price range data points of the general used car data with the Porsche specific data. It can be seen that Porsche sales data typically follows the average trend of general used car data. This can allow Porsche to also look at other companies' sales data when evaluating their own pricing trend. It also allows