Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about

their effect on the dependent variable? (3 marks)

The categorical variables were the most influent in the target variable, as it was showed in the heatmap for correlation. Also, the REF method selected only just categorical variables demonstrating how influential they were to the dependent variable.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

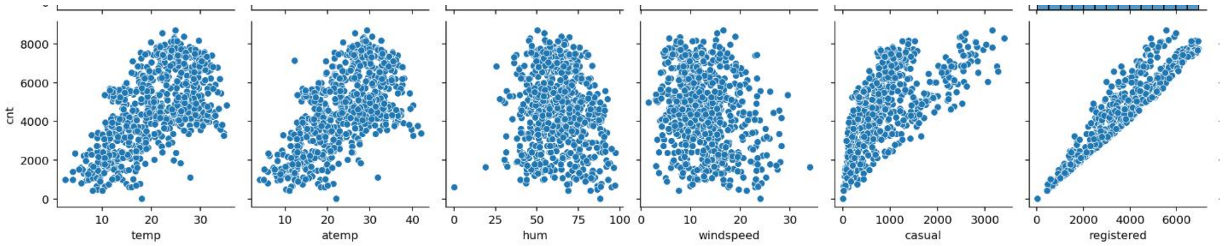
When a dummy variable is created, it creates a column for each of the categories that the variable has, for example for feature season we have four categories. When the dummy features are created, it will create four, one for each category. However, three variables can explain the fourth one, while all the three are 0, that means the fourth one is 1. For that reason, it is only necessary to have 3.

When we use drop_first, one dummy variable is eliminated because it is being explained by the rest of variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable? (1 mark)

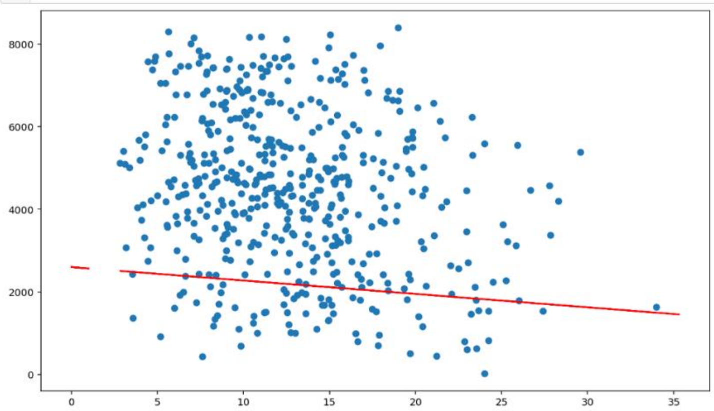
The variable that has the highest correlation with cnt is registered

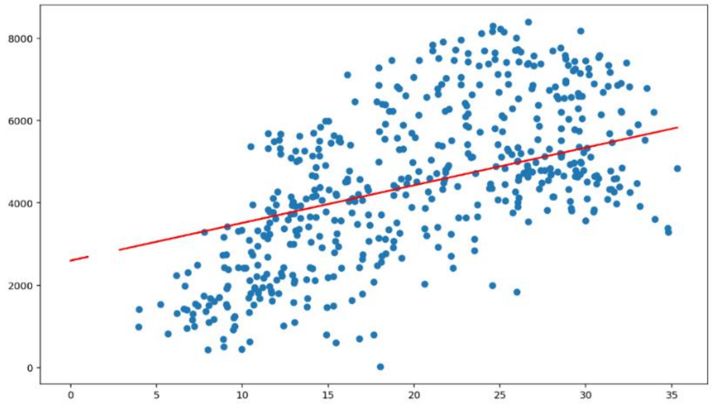


4. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

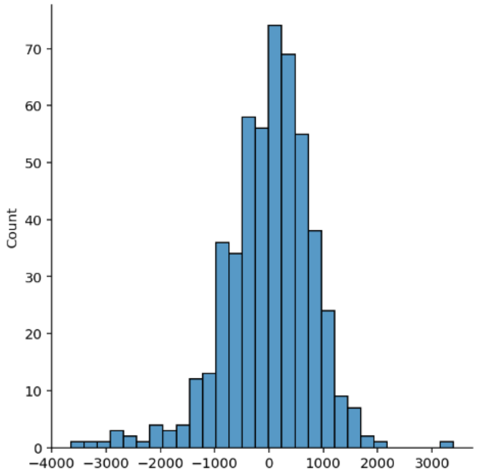
I validate the model by plotting the two numerical variables windspeed and temp against the target variable. One graph was created for each of the features because the training set contained 8 features and cannot be plotted using one single graph because it will be in the nineth dimension. The result of the graphs showed a linear trend.



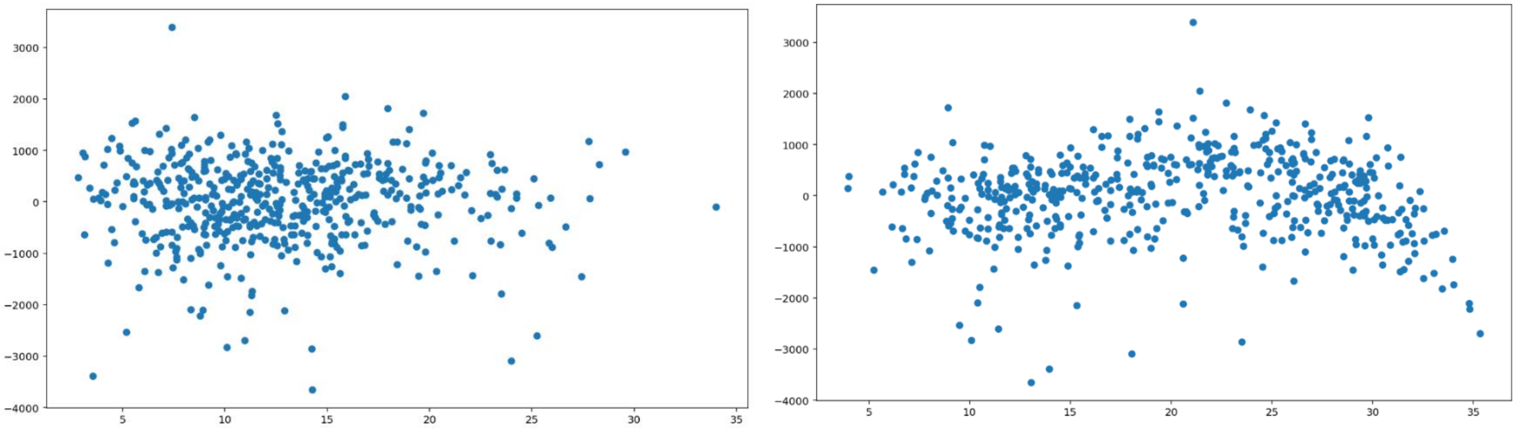


Interr

Also plotting the residuals to verify they follow a normal curve.



Finally, plotting the residuals against the two numerical variables windspeed and temp and verifying that there is not patter in the graph.



5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

For the best model the variables that contributed better to the target were temp, 2019 and spring (variable resulted from the dummy variable creation of yr and season)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

It is a method to predict a target variable (y) by other variables called independent (x). These independent variables should follow a linear relationship with the target. It means

$$y = \beta_0 + \sum \beta_i x_i$$

Where the best beta coefficients are found by the model using a gradient descent method. Then, the model is evaluated with R square score where closer to 1 is better because that indicates that the data is being explain by the model. After, finding the best model or in other words the best coefficients, predictions are performed.

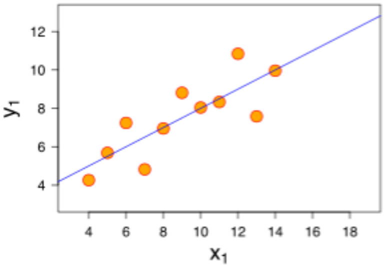
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum.

2. Explain the Anscombe's quartet in detail. (3 marks)

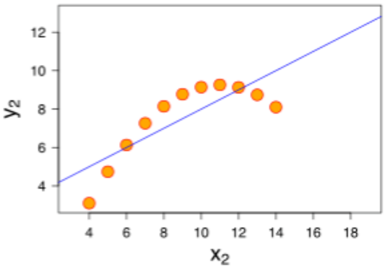
The Anscombe's quartet received its name from its discoverer Francis Anscombe in 1973, who demonstrate the importance of graphing data while being analyzed and the effect that outliers and other influential observations has on statistical properties.

The Anscombe's quartet consists of four data set that have almost identical simple descriptive statistics. However, they have very different when graphed. Each of the set contain eleven points as follow when they are plotted a scatter graph:

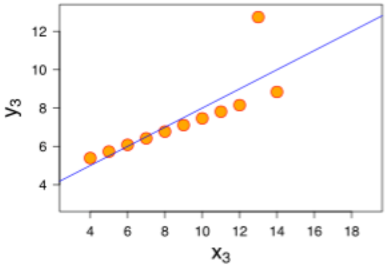
The first looks like a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.



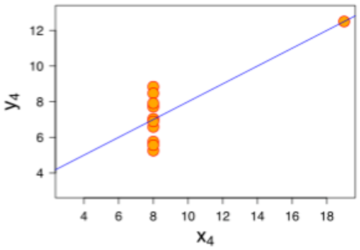
The second it has a relationship which is not linear, and the Pearson correlation coefficient is not relevant.



The third has a linear relationship with the difference that it has one outlier that influence enough to lower the correlation coefficient from 1 to 0.816



The fourth has a single point enough to produce a high correlation coefficient even when the rest of the point do not show a relationship between the variables.



This quartet is relevant because illustrate the importance to always look at the graphs and not only evaluate the basic statistics properties

3. What is Pearson's R? (3 marks)

The Pearson correlation coefficient (*r*) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling? (3 marks)

In simple words scaling is putting all the values into the same range. It is performed to scale features and avoid bias for algorithm considering distances. Also, to avoid extreme coefficients in the Linear Regression model. Two of the more common methods to scaling is normalizing and standardizing.

Normalizing scaling is performed by mapping the point into [0,1] interval. The form to achieve this is by subtracting the minim value to each point and divided by the difference between the maximum value and the minimum.

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardizing scaling is performed by mapping the point into a normal curve with mean equal to 0 and standard deviation of 1, which centralize the data. This is achieved by subtracting the mean of the distribution and divided it by its standard deviation.

$$z = \frac{x - \mu}{\sigma}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

To answer this question let's refer to VIF formula which is:

$$VIF = \frac{1}{1 - R^2}$$

VIF is infinity when the denominator is zero that means $1 - R^2 = 0 \implies R^2 = 1$ i.e. The correlation coefficient is 1 or -1 which means a perfect correlation between two variables. For that reason, if two variables are positive or negative correlated, then, the VIF will be infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

A quantile-quantile (Q-Q) plot is a graphical technique for determining if two data set come from populations with a common distribution. It helps to answer question like:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?

It is important when there are two data samples, and it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.

Bibliographic Resources

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

https://www.baeldung.com/cs/normalization-vs-standardization

https://www.programsbuzz.com/interview-question/you-might-have-observed-sometimes-value-vif-infinite-why-does-happen

https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm#:~:text=The%20quantile%2Dquantile%20(q%2Dq),of%20the%20second%20data%20set.