Assignment-based Subjective Questions

By Chinmay Sahu

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: I have done the categorical variables using boxplot. Below are the points which we observed during visualization:

1. Almost 10% of the bike booking were happening in the months 5,6,7,8,9 & 10 with a median of over 4000 booking per month. The trend increased in the beginning of the year until the middle of the year, when it began to decline as we approached the end of the year. This indicates “mnth“ has some trend for bookings and can be a good predictor for the dependent variable.

2. Almost 64% of the bike booking were happening during ‘weathersit1 with a median of close to 5000 booking. This was followed by weathersit2 with almost 34% of total booking. This indicates, “weathersit” does show some trend towards the bike bookings can be a good predictor for the dependent variable.

3. “Weekday” variable shows very close trend (approx. 14.3% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

4. Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

5. Almost 97.12% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

6. Almost 69% of the bike booking were happening in ‘workingday’ with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable.

Q2. Why is it important to use **drop_first=True** during dummy variable creation?

Answer: drop_first = True is useful because it reduces the extra column created during dummy variable creation. As a result, it reduces the correlations formed between dummy variables. *Drop_first = bool, default False*, indicates whether to extract k-1 dummies from k categorical levels by removing the first level.

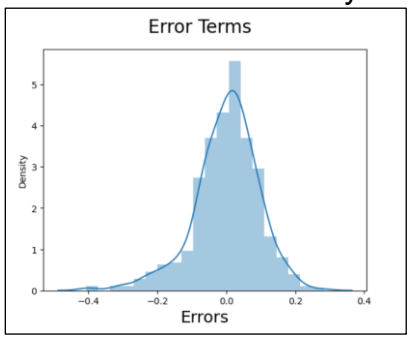
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: So, before model building and training, the pair plot shows highest correlation for " registered" variable having correlation **0.95**. But we are not using casual and registered in our pre-processed training data for model training. casual + registered = cnt. This might leak out the crucial information and model might get overfit. So, excluding these two variables **atemp** & **temp** is having highest correlation with target variable **cn**t. As per the correlation heatmap, correlation coefficient between **atemp** and **cnt** is **0.63**. And correlation coefficient between **temp** and **cnt** is **0.63**.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: I validated the Linear Regression Model assumptions using the five assumptions listed below.

1. Normality of error term - Error terms should be normally distributed.

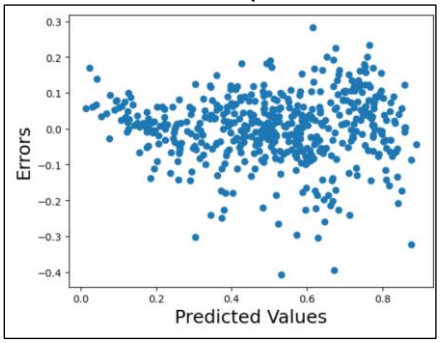


2. Multicollinearity check - There should be insignificant multicollinearity among variables. This taken care as the VIF (Variance Inflation Factor) of all the variables are below 5.

3. Linear relationship validation - Linearity should be visible among variables.

   This is happening because all the predictor variables are statistically significant (p-values are less than 0.05). Also, R-Squared value on training set is 0.823 and adjusted R-Squared value on training set is 0.820. This means that variance in data is being explained by all these predictor variables.

4. Homoscedasticity - There should be no visible pattern in residual values.



   From the above graph obtained there is no clear pattern visible & it is forming cloud of data points which is a good indication.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: Top 3 features significantly contributing towards demand of shared bikes are:

Variable	Coefficient
temp	0.5749
Weathersit_3	- 0.3094
season_4	0.1332

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a supervised machine learning method and a statistical model that investigates the linear relationship between a dependent variable and a set of independent variables. When there is a linear relationship between variables, it means that when the value of one or more independent variables changes (increases or decreases), the value of the dependent variable changes as well (increase or decrease).

The following equation can be used to represent the relationship mathematically.

$$Y = mx + c$$

Here, **Y** is the dependent variable we are trying to predict.
x is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect x has on Y.
c is a constant, known as the Y-intercept. If x = 0, Y would be equal to c.

Assumption for Linear Regression Model are:

1) Linear regression is a powerful tool for understanding and predicting the behaviour of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.
2) Linearity: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.
3) Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.

4) Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.
5) Normality: The errors in the model are normally distributed.
6) No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

Q2. Explain the Anscombe's quartet in detail.

Answer: Before analysing our data and building our model, we must first plot the data set. Anscombe's quartet shows us why.

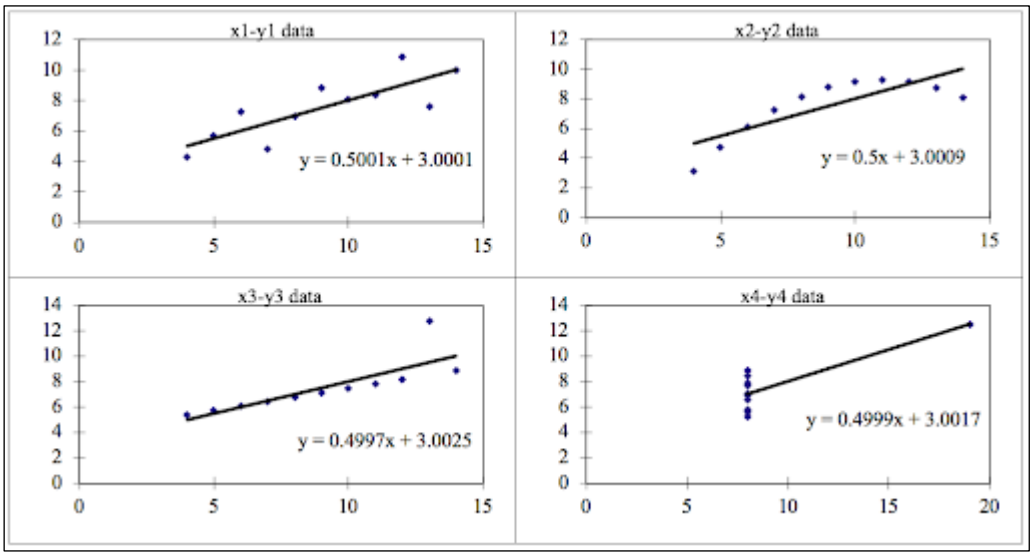
Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

The statistical information for these four data sets is approximately similar. We can compute them as follows:

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



Anscombe's quartet four datasets:

Data Set 1: fits the linear regression model pretty well.

Data Set 2: cannot fit the linear regression model because the data is non-linear.

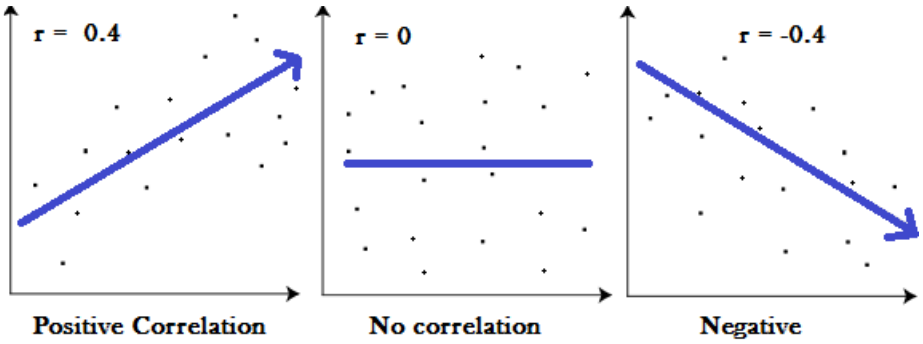
Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

As we can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

Q3. What is Pearson's R?

Answer: Pearson's R or correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.



1) A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
2) A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decrease in (almost) perfect correlation with speed.
3) Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-.95| = .95, which has a stronger relationship than .55.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Feature scaling is a technique for standardising the independent features present in data within a specific range. It is used during data pre-processing to deal with highly varying magnitudes, values, or units. If feature scaling is not performed, a machine learning algorithm will tend to weight greater values as higher and consider smaller values as lower, regardless of the unit of measurement.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Sl No.	**Normalized scaling**	**Standardized scaling**
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$ Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.	$x' = \frac{x - \bar{x}}{\sigma}$ Here, σ is the standard deviation of the feature vector, and x̄ is the average of the feature vector.
3	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
4	Scales values between [0, 1] or [-1,1].	It is not bounded to a certain range.
5	It is really affected by outliers.	It is much less affected by outliers.

6	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: VIF = infinity if there is perfect correlation. A high VIF value indicates that there is a relationship between the variables. If the VIF is 4, it means that multicollinearity has inflated the variance of the model coefficient by a factor of four.

When the value of VIF is infinite, the correlation between two independent variables is perfect. In the case of perfect correlation, R-squared (R2) = 1, resulting in 1/(1-R2) infinity. To address this, we must remove one of the variables from the dataset that is causing the perfect multicollinearity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

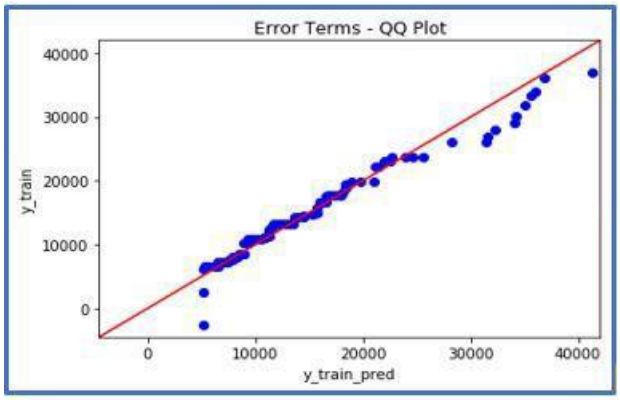
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Interpretation:

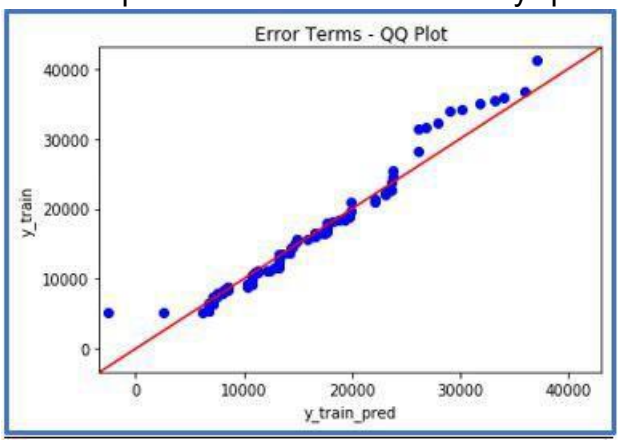
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

1) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
2) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



3) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



4) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis.

statsmodels.api provide **qqplot** and **qqplot_2samples** to plot Q-Q graph for single and two different data sets respectively.