Subjective Questions

Aaron Mathew Alex

Assignment Subjective Questions

Q1.

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

Ans.

The dataset had the following categorical variables:

- season
- yr
- mnth
- holiday
- weekday
- workingday
- weathersit

Below is the analysis for each of them:

season

- Fall experieces the highest percentage of bookings at 32% with a median value of above 5000 bookings (for 2 years)
- Summer and Winter follow with 28% and 25% of the total bookings respectively.
- 2019 had a higher number of bookings as compared to 2018. This suggests that yr also can be a good predictor for the dependant variable.
- This suggests that season can be a good predictor of the dependant variable

mnth

- The preiod from May-October witnessed the highest number of bookings. Each month of this period contributes to 10% or more of the total bookings. Their median values are also above 4000 (for 2 years)
- This suggests that mnth could be a good predictor for the dependant variable

holiday

- 97% of the bike rentals have occured during working days. This indicates that most of the rentals occur during working days.
- This definitely suggests that holiday is a good predictor for the dependant variable.

weekday

- The perc_sum for all the weekdays lie in a similar range (13.5% - 14.8%) and all their medians lie between 4000 and 5000.
- This could indicate that weekday could have less or no influence on the dependant variable.

workingday

- 68% of all the rentals occured on working day.
- Both working and non-working days had a median close to 5000 bookings (for 2 years)
- This suggests that workingday could be a good predictor of the dependant variable

weathersit

- 68% of all the bookings occured on clear days with a median value of almost 5000 bookings.
- This is followed by mist days accounting for 30% of all bookings with a median value of 4000 bookings.
- This suggeststst that weathersit is a good predictor for the dependant variable

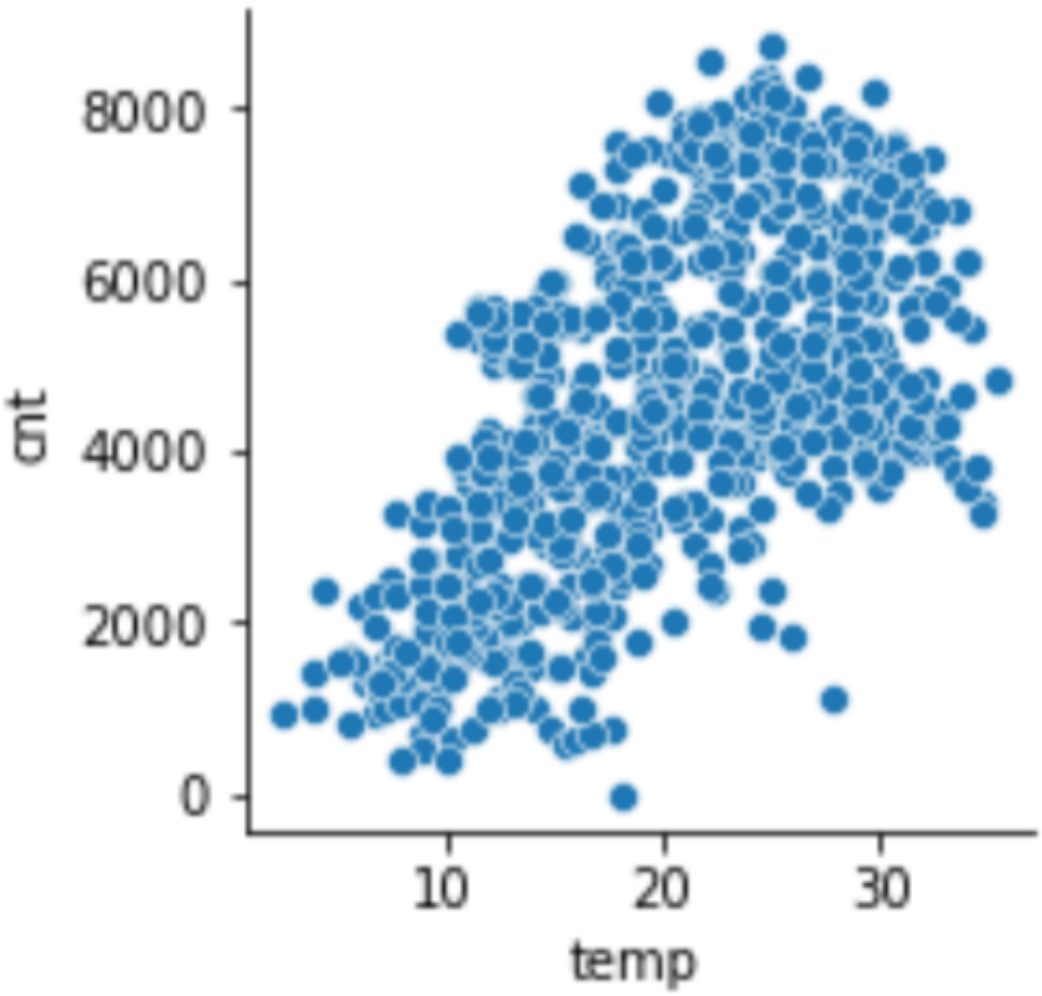
Q2.
Why is it important to use **drop_first=True** during dummy variable creation?

Ans.
If we do nor use drop_first=True we fall into, what is called, the Dummy Variable Trap. The Dummy Variable trap occurs when two or more dummy variables created by one-hot coding are highly correlated (i.e multi-collinearity exists)
For example: If we have a cateogorical variable - Gender with two categories - Male and Female. On creating Dummy Variables, we get X_{male} & X_{female}
Because a 1 in Xmale column would mean a 0 in Xfemale column; we have :

$$X_{male} = 1 - X_{female}$$

This results in two dummy variables that are multi-collinear, and so the dummy variable trap may occur in regression analysis. To overcome the Dummy variable Trap, we drop one of the columns created when the categorical variable were converted to dummy variables by one-hot encoding. This can be done because the dummy variables include redundant information.

Q3.
Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans.
Looking at the pair-plot, **temp** has the highest correlation with our target variable - cnt. Below is the scatter plot between cnt and temp.



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

Ans.
There are 4 assumptions of Linear Regression that I need to validate. They are the following:

1. To check if there is a Linear Relationship between X and y
2. To check if Error terms are Normally Distributed
3. Independance of Error Terms
4. Homoscedasticity - Error terms have constant variance at different values of X

1.
To check if there is a Linear Relationship between X and y
In multiple regression models, nonlinearity or nonadditivity may be revealed by systematic patterns in plots of the residuals versus individual independent variables. The absence of such features affirms the linear relationship between X and y.

I plotted scatter plots of residuals vs individual independant variables and saw no systematic patterns in the plots.

Hence, there is a Linear Relationship between X and y.

2.

To check if Error terms are Normally Distributed

The best test for normally distributed errors is a normal probability plot or normal quantile plot of the residuals. These are plots of the fractiles of error distribution versus the fractiles of a normal distribution having the same mean and variance. If the distribution is normal, the points on such a plot should fall close to the diagonal reference line.

I observed that:

- The error terms were normally distrubted on the probability plot
- The points plotted on the Q-Q plot did fall on the diagonal reference line.
-

Hence, error terms are normally distributed

3.

Independance of Error Terms

The best test for serial correlation is to look at a residual time series plot (residuals vs. row number) and a table or plot of residual autocorrelations. The Durbin-Watson statistic provides a test for significant residual autocorrelation at lag 1: the DW stat is approximately equal to 2(1-a) where a is the lag-1 residual autocorrelation, so ideally it should be close to 2.0--say, between 1.4 and 2.6 for a sample size of 50. The closer it is 2, the less the auto-correlation between the various variables.

My model has a DW value of 2.0467.
There is almost no auto-correlation. Therefore this suggests that the residuals are indeed independant of each other.

Hence, error terms are independant of each other

4.
Homoscedasticity - Error terms have constant variance at different values of X
A scatterplot of residuals versus predicted values is good way to check for homoscedasticity. There should be no clear pattern in the distribution; if there is a cone-shaped pattern the data is heteroscedastic.

I plotted the above plot and saw that there is no clear pattern in the distribution. This affirrms that the error terms have constant variance at different values of X.

Hence, error terms are homoscedastic

Q5.
Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans.
This is the final model we have arrived at:

$$
\begin{aligned}
cnt = {} & 0.239423 + \\
& (0.234878*yr) + \\
& (-0.090839*holiday) + \\
& (0.423611*temp) + \\
& (-0.159255*windspeed) + \\
& (-0.059974*spring) + \\
& (0.048122*summer) + \\
& (0.100637*winter) + \\
& (-0.045116*dec) + \\
& (-0.052486*jan) + \\
& (-0.041664*nov) + \\
& (0.081832*sep) + \\
& (-0.293808*lrainsnow) + \\
& (-0.080015*misty)
\end{aligned}
$$

Based on this model, the 3 top features are:

1. Temperature (Positive Influence)
2. Light rain & snow (Negative Influence)
3. Year (Positive Influence)

General Subjective Questions

Q1.
Explain the linear regression algorithm in detail.

Ans.
Machine Learning algorithms are of three types - Supervised Learning, Unsupervised Learning and Reinforcement Learning. Regression is a type of supervised learning.
Regression models a target prediction value based on independant variables.
Linear Reagression help predict a dependant variables (y) based given independant variables (X). It tries to find a linear relationship between X and y.
Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent (X) and dependent (y) variable.
Based on the given data, we try to model a linear equation as below:

$$y = \beta_0 + \beta_1 \times x$$

Our motive is to find the best values for b0 and b1 . We try to derive the equation of the line that fits the data in the best way. The way we do this is by using something called Residuals.

What are Residuals?

The residual for a given X value, is the difference between the value predicted by the line and the acutal Y value.

$$e_i = y_i - y_{pred}$$

This value can either be positive or negative. Therefore, adding up these residuals would give us a false representation. We will therefore square them.

$$e_i^2 = (y_i - y_{pred})^2$$

Summing up these values for all values of X gives us the Residual Sum of Squares.

$$RSS = \sum_{i=1}^{n} e_i^2$$

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_0 X_i)^2$$

$We\ should\ ideally\ select\ \beta_0\ and\ \beta_1\ such\ that\ RSS\ is\ minimum.$

Cost Function

The cost function helps us to figure out the best possible values for beta0 and beta1 which would provide the best fit line for the data points. Therefore, we convert this problem into a minimization problem where we would like to minimize the error.

$$minimize \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{pred})^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{pred})^2$$

We choose the above function to minimize. The difference between the predicted values and ground truth measures the error difference. We square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error (MSE) function.

Gradient Descent

To update beta0 and beta1 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent.
The idea is that we start with some values for beta0 and beta1 and then we change these values iteratively to reduce the cost. Gradient descent helps arrive at the best values. Gradient Descent points in the direction of steepest descent from any given point. Starting at any point on the curve of our Cost Function, we take our first step in the direction specified by the negative gradient. Next we recalculate the negative gradient (passing in the coordinates of our new point) and take another step in the direction it specifies. We continue this process iteratively until we get to the bottom of our graph, or to a point where we can no longer move downhill–a local minimum.

Learning Rate

The size of these steps is called the learning rate. With a high learning rate we can cover more ground each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing. With a very low learning rate, we can confidently move in the direction of the negative gradient since we are recalculating it so frequently. A low learning rate is more precise, but calculating the gradient is time-consuming, so it will take us a very long time to get to the bottom.

Strength of Linear Regression

Once we have arrived at the best values for our coeffecients, we need to know how well the best fit line represents the scatter plot. We can't use RSS for this as it is absolute; which means that the RSS value will change if we change the units of either of the variables. We use a measure called **Rsquare**.

$$R^2 = 1 - \frac{RSS}{TSS} \quad Where\, TSS = \sum_{i=1}^{n}(y\_ - y_i)^2$$

Rsquared can take a value between 0 and 1.
1 symbolizes the best fit while 0 symbolizes the worst fit.
Suppose our model has Rsquared value of 0.83. We interpret it the following way:
Our model is able to explain 83% of the all the variance in the data.

Q2.
Explain Anscombe's quartet in detail.

Ans.
In 1973, the English statistician Francis Ascombe constructed a group of four datasets to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.
Here are the four datasets:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

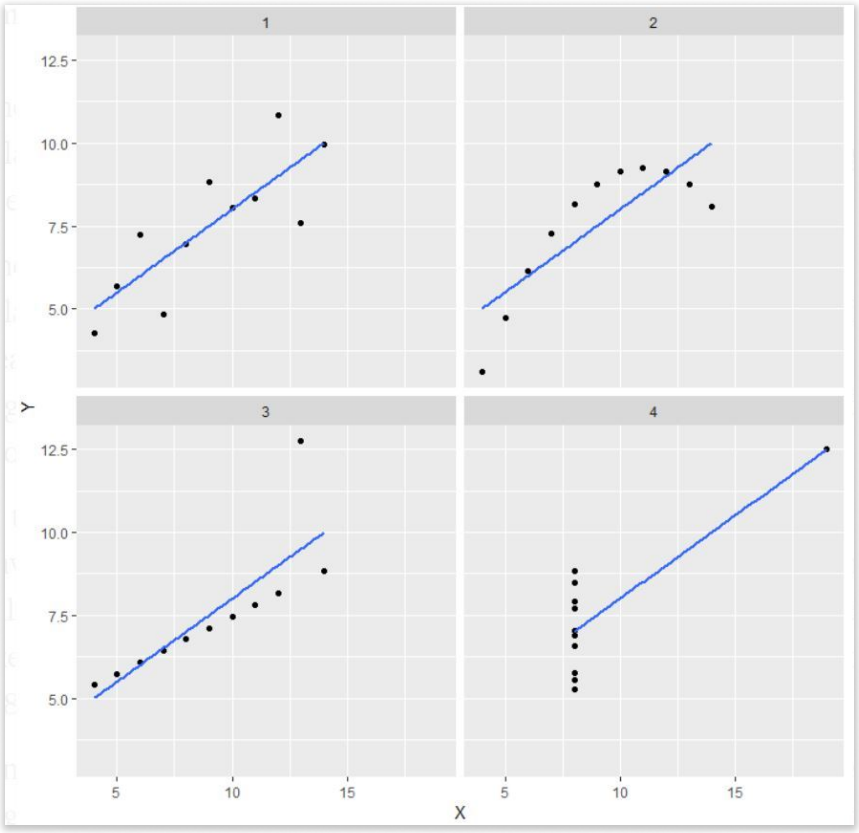
On analyzing them using only descriptive statistics, we find the mean, standard deviation and correlation between x & y.

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

Let us however plot these models on a scatter plot.



We can now describe the 4 datasets as:
Dataset 1: This fits the linear regression model pretty well.

Dataset 2: This could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: Shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: Shows the outliers involved in the dataset which cannot be handled by linear regression model

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities shown in the descriptive statistics.

References:

- Importance of Data Visualization — Anscombe's Quartet Way (Sparsh Gupta)
- Anscombe's Quartet (GeeksforGeeks)

Q3.
What is Pearson's R?

Ans.
Correlation Coeffecients are used to measure how strong a relationship is between two variables. Of the several different types of correlation coeffecients, the most popular one is Pearson's. Here is the formula for calculating the same.

Formula: $$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

The formula returns a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

However, the Pearson's R correlation method has some limitations as well.

- Correlation is not and cannot be taken to imply causation. Even if there is a very strong association between two variables we cannot assume that one causes the other.
- The method is not able to tell the difference between dependent and independent variables. For example, if you are trying to find the correlation between age and a person's income, you might find a positive correlation. However, you would get the same result if you switched the variables around.
- Correlation does not allow us to go beyond the data that is given. For example suppose it was found that there was an association between time spent on homework (1/2 hour to 3 hours) and number of G.C.S.E. passes (1 to 6). It would not be legitimate to infer from this that spending 6 hours on homework would be likely to generate 12 G.C.S.E. passes.

References:

- Correlation Coeffecients (StatisticsHowTo)
- Correlation Definitions, Examples & Interpretation (Simply Psychology)

Q4.
1. What is scaling?
2. Why is scaling performed?
3. What is the difference between normalized scaling and standardised scaling?

Ans.
1. Scaling or Feature Scaling as it is more commonly known, is a method to standardize the independant variables present in the data and bring them within a fixed range. This is usually done in the data pre-processing stage to deal with data that is in highly varying magnitudes or units.

2. A dataset is just a bunch of numbers to a ML algorithm.
If there are huge differences in the range of values we have, the algorithm makes the assumption that the larger numbers have some sort of superiority. This is attributed to ML algorithms using Euclidean distance to calculate the distance between two data points in their computation.
If left alone, these algorithms only take in the magnitude of the features, neglecting the units.
The results would vary greatly between different units, 2km and 2000m. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To suppress this effect, we need to bring all features to the same level of magnitudes.
This can be achieved by scaling.

3.

Normalized/Min-Max Scaling:

This brings all of the data in the range of 0 - 1.

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

Standardised Scaling:

This brings all the data into a standard normal distribution which has mean as zero and standard deviation as one.

Formula: $$x = \frac{x - mean(x)}{sd(x)}$$

References:

- Feature Scaling for Machine Learning - (Aniruddha Bhandari)
- Feature Scaling- Why it is required? (Rahul Saini)
- All about Feature Scaling (Baijayanta Roy)

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

Ans.
An infinite VIF value indicates that the corresponding variable can be expressed exactly as a linear combination of other variables; which also show an infinite VIF value as well.
This depicts a perfect correlation between two independant variables. . In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans.
The Q-Q (quantile-quantile) plot is a graphical technique for determinig if two data sets come from populations with a common distribution.