Assignment Based 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: Based on Exploratory data analysis on the dataset, the following can be said:

a) Season: Highest number of bikes were rented in fall and summer season
b) Year: More bikes were rented in 2019 than 2018
c) Month: Count of rented bikes increases from Jan to Mid of year, reaches max in Sep and then again decreases
d) Holiday: Less bikes are rented on holiday
e) Weather: Maximum number of bikes are rented on days when weather is Clear and minimum are rented in rainy weather.
f) No clear trend of no of bikes with weekday or workingday.

==========

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

Ans: We use drop_first=True in dummy variable creation for feeding the categorical data to the model. It is important to use for following reasons.

a) To drop the redundant variable that is created.
   If a column has n categories, we only need n-1 dummy variables to represent all the information.
b) To reduce the multicollinearity in the data because the dummy variables will be highly correlated with each other.
   Example: To represent a column with categories – Heads and Tails.

<table>
<tr><th>WITHOUT drop_first=True</th><th>WITH drop_first=True</th></tr>
<tr><td>Two columns are created: Heads and Tails. If it is heads, mark as 1 under Heads and 0 under tails

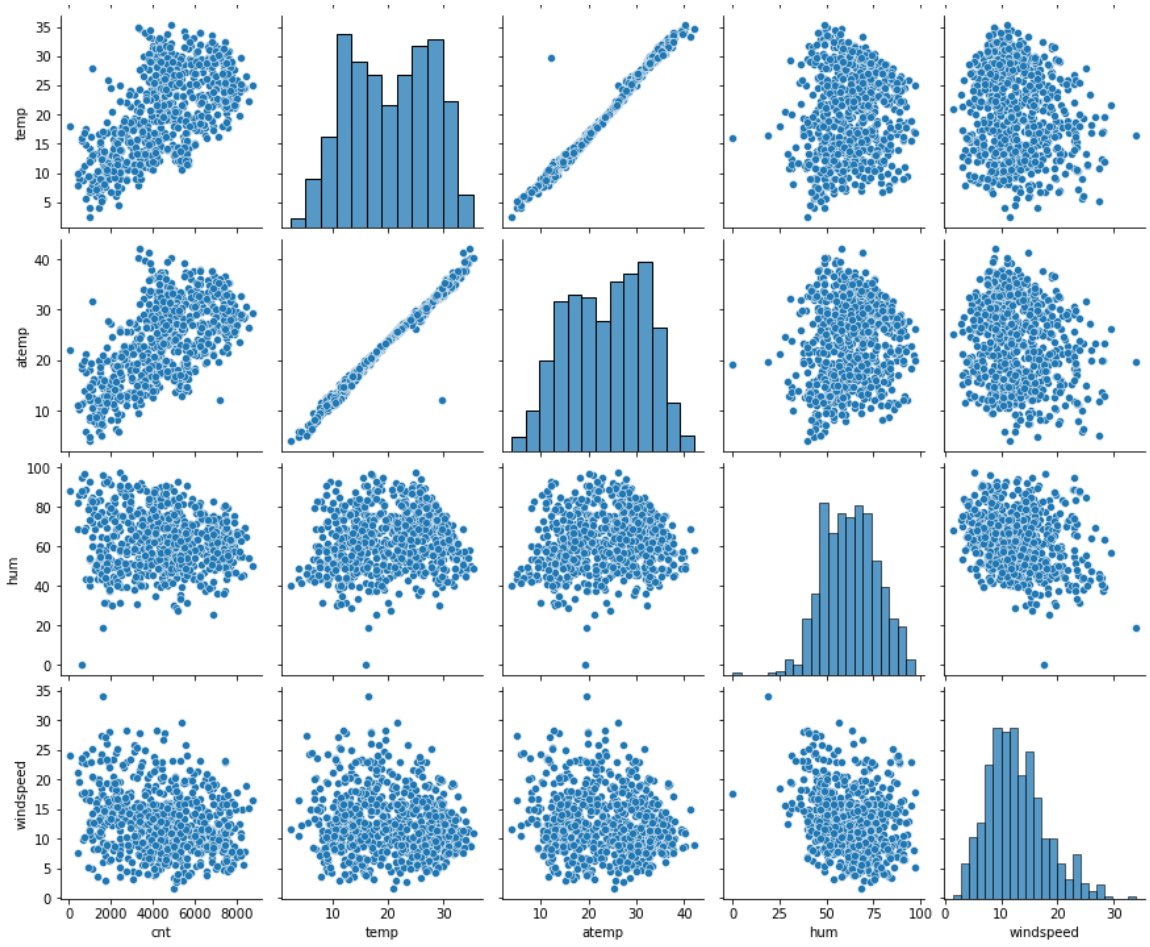
<table><tr><th>Heads</th><th>Tails</th></tr><tr><td>1</td><td>0</td></tr><tr><td>0</td><td>1</td></tr><tr><td>0</td><td>1</td></tr><tr><td>1</td><td>0</td></tr><tr><td>1</td><td>0</td></tr></table></td><td>Only 1 column can represent the data. If 0, then heads Otherwise tails.

<table><tr><th>Tails</th></tr><tr><td>0</td></tr><tr><td>1</td></tr><tr><td>1</td></tr><tr><td>0</td></tr><tr><td>0</td></tr></table></td></tr>
</table>

==========

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 (*image attached*), **temp** has the highest correlation with target variable 'cnt'.

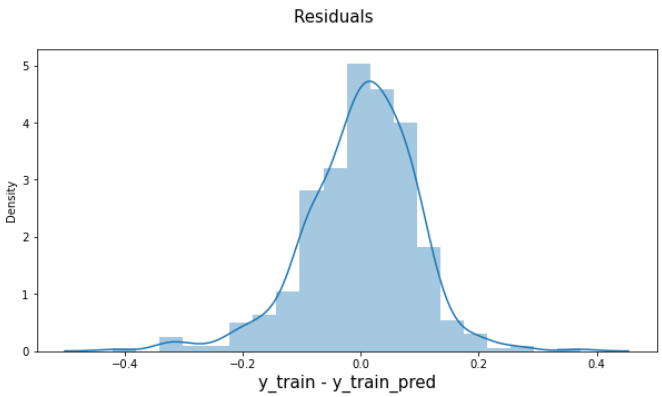


==========

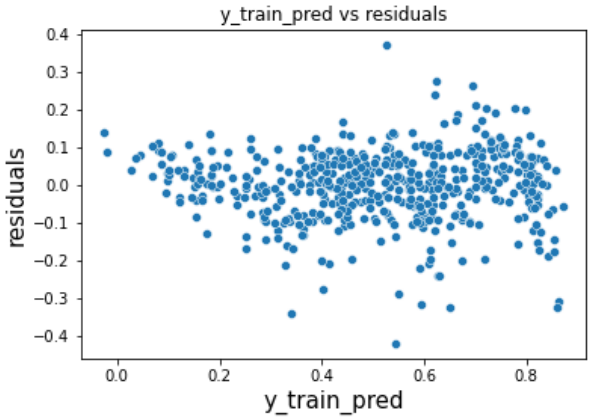
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans:

a) To validate that error terms distribution is normal and centred around 0, I plotted the histogram (dist plot) for error terms.
So, the distribution is bell shaped <u>normal distribution, centred around 0</u>.



b) To validate that error terms, have no pattern with y or x, and are completely random, I created a scatter plot and found no specific patterns. Since, the points are completely random, the error terms have constant variance (homoscedasticity)



==========

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

Ans:

The top 3 features are:

a) temp (Positively correlated)
b) Rainy (negatively correlated)
c) yr (Positively correlated)

	Cofficient	Sign
temp	0.332015	+
Rainy	0.282162	-
yr	0.237313	+
windspeed	0.156730	-
spring	0.133961	-
holiday	0.090153	-
Cloudy	0.076706	-
Sep	0.069269	+
Jan	0.056987	-

==========

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a supervised learning method that uses the past data with labels to predict a continuous/ numeric variable. E.g.: How many likes will this meme get?

Linear Regression is of 2 types:

1. Simple Linear Regression:

In this case, we build a model for only 1 independent variable.

The basic idea of linear regression is to *fit a line that passes through the maximum number of points in the data*.

If y in the dependent variable and x is the independent variable, the value of y for any x is given as:

The equation y = Bo + B1x where B1 is the slope of the graph and Bo is the y-intercept (value of y when x is zero).
B1 signifies relationship between x and y).

If the value of x is increased by 1 unit, the value of y will increase by B1.

If B1 is +ve, means y increases with increase in x. If m is -ve, means y decreases with increase in x.

Determining best fit line

For this we use residuals. Residual is the error associated with each data point and is given as difference between actual value of y and predicted value of y.

Residual = y (actual) – y(pred)

The best fit line will be the one for which sum of squares of these error terms/ residuals (RSS) is minimum. This method is called Ordinary Least Squares Methos (OLS).

RSS is a relative measure and will change with units.

Other measure is TSS. TSS is the benchmark where the model will predict the average value of y for every value of x. So TSS is given as sum of square of difference between actual value of y and average value of y.

Strength of the model

An absolute measure to access goodness of fit is called R squared. It is given by :

R squared = 1 – (RSS/TSS)

Here RSS is expected to be less than TSS. Since we are talking about errors. So errors should be minimised.

An R -squared value ranges between 0 and 1. More is the R squared, more variance is explained by the model and better is the linear fit

RSS	TSS	R squared	Interpretation
0	10	1	There are no errors in data and it is an ideal scenario. (May be possible when model is overfit). Model is able to explain all variance in data
6	10	0.4	Model is able to explain 40% variance in data. Our model is 40% better than the model that simply predicts the average
10	10	0	Model is worst and is not able to explain any variance in y

Cost function

Machine learning problems define a cost function that has to be minimized or maximized. In our case, the cost function is RSS and we need to minimize it and find the best values for Bo and B1.

Hypothesis

After we get the Bo and B1 from the model, we need to test whether these coefficients are significant or not. To do this, hypothesis testing is used with alpha = 0.05

Null Hypothesis: B1 =0

Alternate Hypothesis: B1! = 0

If the p-value for this comes out to be less than 0.05, we reject null hypothesis and conclude the B1 has a significant impact on y. If the p-value for this comes out to be more than 0.05, we fail to reject null hypothesis and conclude the B1 is insignificant.

Assumptions

a) There must be some linear relationship between X and y
b) Since we are trying to generalize from a sample to a population so bring in uncertainty in form of error terms that are associated with each X. Error terms are normally distributed.
c) They are independent of each other.
d) Mean of these error terms is 0
e) Errors have a constant variance throughout the values of x (homoscedasticity)

2. Multiple Linear Regression

In this case, we build a model for more than 1 independent variable i.e., multiple predictors.

The equation for this is given as y = Bo + B1x1 + B2x2 +…….+ Bnxn

When x1 is increased by 1 unit, y increases by B1 provided all other features are constant.

Issues to deal with in multiple regression involve:

a) Multicollinearity
b) Feature scaling
c) Model Assessment and Comparison
d) Feature selection

Multicollinearity

When we have one or more variables that are highly correlated to each other

a) We cannot separate how each of these is impacting the dependent variable y.

For ex: y = Bo + B1x1 + B2x2

If B1 = 0.1 and B2 = 0.9, we may conclude that since B 1 is less, it is less useful. But in actual, it may be possible that if X1 and X2 are highly correlated, all variance in y is explained by x2.

b) Interpretation

We define y = Bo + B1x1 + B2x2 +……. + Bnxn

When x1 is increased by 1 unit, y increases by B1 provided *all other features are constant*. If the independent variables are highly correlated, they will not be constant.

c) Coefficients may change their signs and p-values will also change.
Earlier if the predictor was significant (p-value < 0.05), on introducing a highly correlated variable, the predictor may become insignificant (p-value > 0.05)

VIF

We should eliminate the highly correlated predictors in the model building using a VIF factor. VIF is calculated for all the predictor variables

VIF (Variance Inflation Factor) = 1 / (1-R2)

If VIF is high, means the predictor variable is highly correlated to other predictors and must be dropped if VIF > 10.

If VIF is low (<5), means these is very less correlation.

Scaling

Scaling is an important step in preparing the data for analysis and it brings all the numeric data within a certain range.

Scaling is performed for the following reasons:

a) To easily *interpret the coefficient* of predictors and understanding the linear relationship between dependent variable y and independent variable(s)

E.g.: If Area is in thousands and number of bedrooms is in digits. The coefficient of area will be less and bedrooms will be more. We can wrongly interpret that number of bedrooms is a strong predictor of Price because of large coefficient. But it is wrong. So, we bring all the features on same scale.

b) To *speed up the back-end conversion performed by Gradient Descent Algorithm* that minimizes the cost function

Scaling can be done in 2 ways:

Diff	Normalized Scaling	Standardized Scaling
Spread	Data is normalized between 0 and 1	Data is normally distributed such that mean is 0 and standard deviation is 1
How	Uses Min and max of a feature to scale	Uses mean and standard deviation to scale
Outliers	Impacted by outliers	Less impacted by outliers
Binary variables	No impact on binary (0,1) variables	Distort the binary (0,1) variables data
Formula	(x – xmin) / (xmax – xmin)	(x – mean) / sd
SciKit Learn class	MinMaxScaler	StandardScaler

Model Assessment and Comparison

We want a model that is simple and at the same time, explains maximum variance in y.

To accomplish this, Adjusted R2 is introduced.

Adjusted R2 = $1 - \frac{(1-R2)\,(N-1)}{N-p-1}$ where N is the number of records, p is number of predictors

Adjusted R2 *penalizes the model based on every additional independent variable added*. If an insignificant variable is added to the model, adjusted R2 will decrease unlike R2 that always stays the same or increases for every additional variable added.

Feature selection

Feature selection can be done using Recursive Feature Elimination (RFE) or manual techniques.

Steps for creating model

1. Understanding and visualizing the data (Exploratory Data Analysis)
2. Data Preparation for modelling including encoding (categorical variables), train-test split, rescaling
3. Training the model
4. Residual analysis

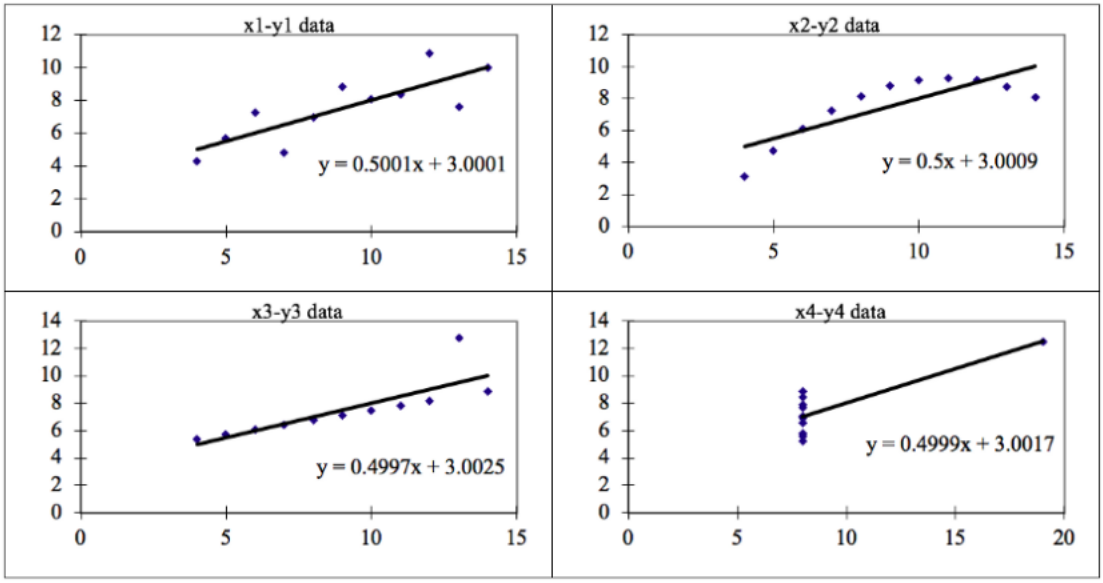
5. Prediction and evaluation on the test set

==========

Q2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet consists of four data sets, having 11 points (x,y) each, having similar descriptive statistics (like mean, variance, correlation coefficients and even same best fit line). One might think since the statistics for these 4 graphs is same, the distribution of data points might also be similar. But, in actual, all 4 data sets have a very different distribution when plotted.

Below are the graphs obtained for the 4 datasets:



There are the four graphs:

Graph 1: The data points are almost symmetric about line of best fit indicating linear relationship between x and y

Graph 2: The data points have a curved pattern indicating non-linear trend

Graph 3: Almost all the points lie on the line of best fit, apart from one outlier

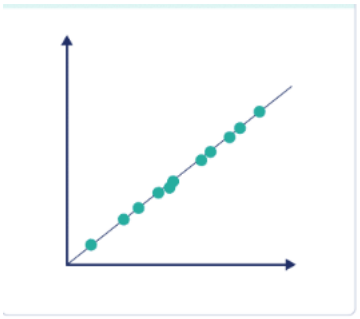
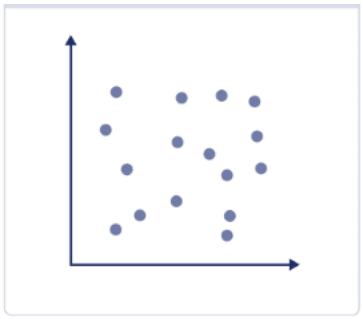
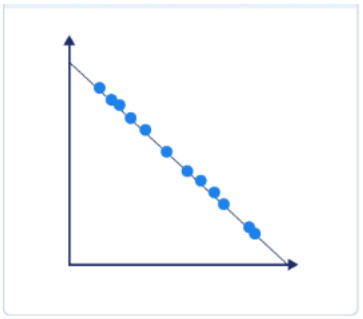
Graph 4: One outlier has drastically affected the graph.

To conclude with, it can be said that statistics alone can fool easily. It is always important to use statistics in conjunction with visualizations / plots so that we don't fail to get the whole picture. Thus, it is very important to visualize the data before drawing any conclusions.

==========

Q3. What is Pearson's R?

Ans: Pearson's R measures the linear relationship between 2 variables. It will always lie between -1 and 1.

Graph 1 indicates strong positive linear relationship between x and y with Pearson's R = 1

Graph 2 indicates no relationship between x and y with Pearson's R = 0

Graph 3 indicates strong negative linear relationship between x and y with Pearson's R = -1

Pearson's R is given by the formula:

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{10}}{\sqrt{(\sum x_i^2 - \frac{(\sum x_i)^2}{10})}\sqrt{(\sum y_i^2 - \frac{(\sum y_i)^2}{10})}}$$

==========

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

Ans: Scaling is an important step in preparing the data for analysis and it brings all the numeric data within a certain range.

Scaling is performed for the following reasons:

a) To easily ***interpret the coefficient*** of predictors and understanding the linear relationship between dependent variable y and independent variable(s)

Eg: If Area is in thousands and number of bedrooms is in digits. The coefficient of area will be less and bedrooms will be more. We can wrongly interpret that number of bedrooms is a strong predictor of Price because of large coefficient. But it is wrong. So we bring all the features on same scale.

b) To ***speed up the back-end conversion performed by Gradient Descent Algorithm*** that minimizes the cost function

Scaling can be done in 2 ways:

Diff	Normalized Scaling	Standardized Scaling
Spread	Data is normalized between 0 and 1	Data is normally distributed such that mean is 0 and standard deviation is 1
How	Uses Min and max of a feature to scale	Uses mean and standard deviation to scale
Outliers	Impacted by outliers	Less impacted by outliers
Binary variables	No impact on binary (0,1) variables	Distort the binary (0,1) variables data
Formula	(x – xmin) / (xmax – xmin)	(x – mean) / sd
SciKit Learn class	MinMaxScaler	StandardScaler

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

Ans: Infinite VIF for a predictor variable A means perfect correlation between

- the predictor A and other independent variable B OR
- between predictor A and linear combination of other predictor variables (B1, B2, …. Bn).

As per the formula, VIF = 1 / (1-R2) where R2 is the R-squared.

When R2 is 1, which means the two independent variables are perfectly correlated (Ideal scenario).

If VIF is infinite, it is always better to drop one among the highly correlated variables, so that the assumption that predictor variables have no or very less correlation among themselves still holds for performing linear regression.

==========

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

Ans: Q-Q plot, also known as Quantile-Quantile plot is created by plotting quantile of one set of data against the other set of data.

Q-Q plot will help to determine the distribution (normal, uniform, etc.) that a set of data follows.

For e.g.: If we have a variable B whose distribution is unknown.

Check whether B follows normal distribution or not?

Step 1: We take an existing normal distribution of any other variable A.

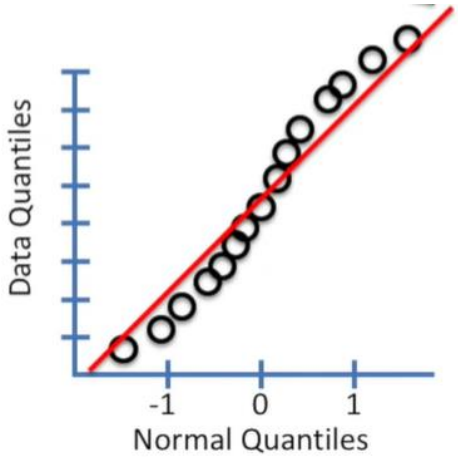
Step 2: We plot the quantiles of both these sets on 2 axes: Quantiles of normal distribution on x-axis and quantiles of unknown data on y-axis.

Step 3: We plot all the points representing the location where quantiles of our dataset intersect with normal quantiles.

Step 4: We check if the points lie on the straight line.

If most of the points fit the line, we can conclude distribution of the dataset is normal.

Else, we conclude distribution of the dataset is not normal.

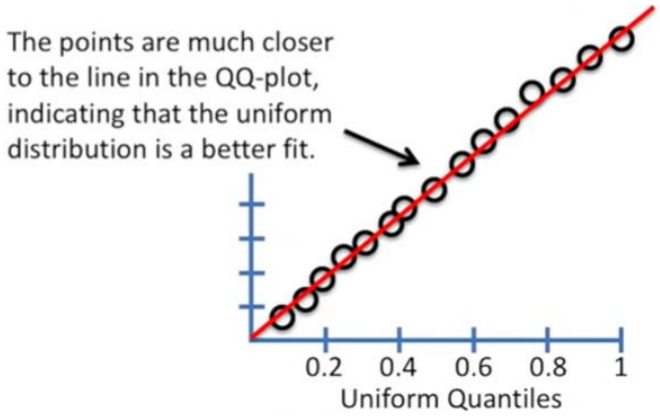


In this case, the fit is not really good. If our data was normally distributed, most of points would have lied on the line. Hence, this distribution is not normal.

We need to compare the quantiles of our dataset with some other distribution quantiles.

Check whether B follows uniform distribution or not?

Using the uniformly distributed quantiles of a set of data A, the same steps mentioned above can be followed.



Here, all the points hug the line very nicely, indicating that B follows uniform distribution.

==========