Assignment-based Subjective Questions

Q 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans :- Below are the observation from the boxplot graph of case study.

1. Season :- In the fall season demand of rent bike is high and low in spring
2. Month(mnth) :- In the September month demand of rent bike is high and low in Jan.
3. Weekday :- In Saturday (weekday) demand of rent bike is high and low in Tuesday.
4. Weathersit :- In Clear_Few_clouds_Partly_cloudy weather demand of rent bike is high and low in Light_Snow_Rain_Thunderstorm_Scattered_clouds.
5. Working day and Holiday :- In normal working day demand of rent bike is high and low in weekends or holiday.

Q 2. Why is it important to use drop_first=True during dummy variable creation?
Ans :- drop_first = True is important in dummy variable creation . There are below reasons.

1. Reduce the number of columns . i.e. if dummy variable created three new column this method will reduce it to n-1 column.
2. It reduce the co-relations among variables.

Why in terms of Linear Regression .

- Get dummy variable divide the categorical level to 0,1 (binary form) if we have three category level like for season :- summer, spring, winter , fall then get_dummy will divide it into as 1,0,0,0 and 0,1,0,0 and 0,0,1,0 and 0,0,0,1 So here we can remove first one so if all three are 0,0,0 then its called summer .

Q 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans :- atemp

Q 4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans :-

1. Linear regression require a relationship between independent variables and dependent variable (Target variable).Now we create a pair plot with y & x variables . (y as Target variable & x as independent variables) . From pair plot you can know the relation between variables and you can know the linearity is present or not.
2. Data clean up :- Check null values, Outliers, not useful variables (No effect on dependent variables, data collinearity . Remove or derived data from them or fill with related values.
3. Correlation matrix .
4. Check categorical data and get_dummies variables and scale the variables .
5. Divide data into test- training data set .
6. First fit & transform model with training data
7. Check R-square , Adj-R Square , P value, and coefficient . R-square should be high enough to find fit lines not overfit lines . P value should be less than 0.05 .

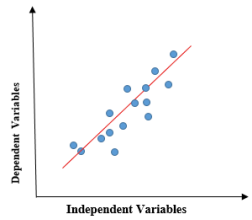
8. Check VIF (Variance inflation factor allow to determine the strength of the correlation between the different independent variables) . VIF value should be less then 5 for good fit line.
9. Now transform the test data with the variables fit for linear line from the training set model . and find the value of r2 score .
10. If both Training and Test r2 score is good high enough for fit line then your model is Good linear regression model .

Q 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans :- Top 3 features contributing towards explaining the demand of the shared bikes.

1. atemp (feeling temperature in Celsius) :- Positive Corelation
2. weathersit (Light_Snow_Rain_Thunderstorm_Scattered_clouds + Mist_Cloudy_Broken_Few_clouds) :- Negative Co-relation
3. yr (Year) :- Postive Corelation

General Subjective Questions

Q 1. Explain the linear regression algorithm in detail.
Ans :- Linear regression is a Machine learning algorithm based on supervised learning. It performs a regression task . It's basic purpose is to predict target values based on independent variables (X). This regression technique find-out a linear relationship between X(input) and y (Output) variables .
"Regression shows a line that passes through all the data points on a target-predictor graph is such a way that the vertical distance between the data points and the regression line is minimum."



Linear regression equation is :

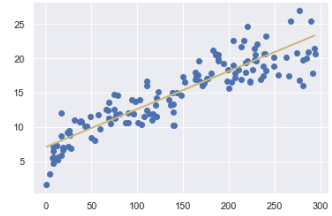
$Y = mx + c$

m (Slobe) , c (Intercept) , x (Independent variables) , y (Target variable)

In linear regression you have to find value of m , c so you can fit the line using y prediction .

1> Draw a pair plot with target & independent variables . So you can know how much it is corelated or you can create heat map.
2> Divide data set into training & test data set
3> Find ordinary least square using sm.OLS method
4> Fit the model with the same linear regression and get the summary
5> From summary fetch the value of m & c coef.
6> Finally fit the line with formula y = mX + c

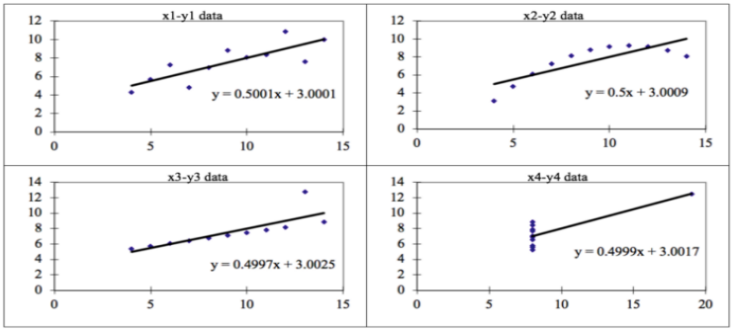
```
In [157]: plt.scatter(X_train,y_train)
          plt.plot(X_train,7.105951 + 0.055313*X_train, 'y')
          plt.show()
```



c = 7.105951, m = 0.055313

Q 2. Explain the Anscombe's quartet in detail.
Ans. Anscombe's quartet is a group of four data sets which is used to visualize data before applying any algorithms to build model. This suggest us the data features must be plotted to see the distribution of sample data that can help us to find the data anomalies like outliers , diversity , data is linear or not etc. Because linear regression is only useful for data which is in linear relationship no other type of data.

When any models are plotted on scatter plot all datasets generate a different kind of plot that is not interpretable by any regression algorithm.



From the model these datasets can be described as :

1. DataGraph1: This means its fits linear regression model pretty well.
2. DataGraph2: This means data is nonlinear. It couldn't fit linear regression.
3. DataGraph3,4 : This means data have outliers which cannot be handled by the linear regression.

Conclusion :-
So Anscombe's quartet is quiet useful to understand the data visualization . So before attempting to interpret and model the data or any algorithms. We have to first visualize the data set in order to build a good fit model .

Q 3. What is Pearson's R?
Ans :- The Pearson correlation coefficient also referred as Pearson's R or bivariate correlation. It is a measure of linear correlation between two data sets. It is a numerical summary of the strength of the linear association between two variables. If its go up and down together its called positive correlation and if it go up and down opposite sites called negative correlation.

The Pearson's correlation coefficient varies between – 1 and +1 .

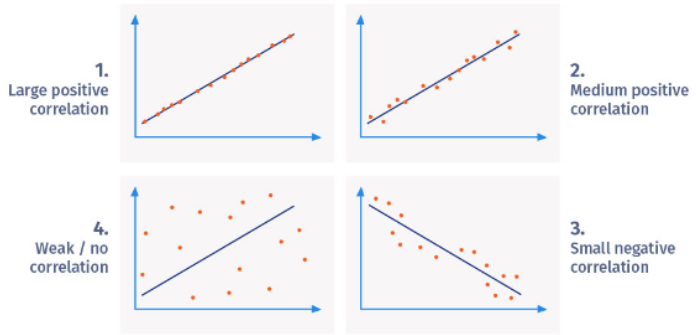
Pearson r formula =

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

Where
r = correlation coefficient
xi = value of x variables
xbar = mean of x variables value
yi = value of y variables
ybar = mean of y variables value

What we can understand from r values

r = 1 (Linearly associated with high positive correlation)
r = -1 (Linearly associated with high negative correlation)
r = 0 (No Linearly associated)
r >0 <5 = (Week Linearly associated)
r > 5 < 8 = (Moderate Linearly associated)
r > 8 = (Strong Linearly associated)



Q 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: Scaling is a data Pre-processing steps . It is used to normalize data in to the particular range of the independent variable. Like if we have set of data with very high and low values with different unit . Then if we apply any Algorithm it can understand only values not units of value so result will not be always confident result . It can mislead us. So to overcome this problem we used scaling that used to bring all the variables in the same magnitude.

Most important point is this normalization only effect coefficient no other result parameter like R-Square, Adj-R-Square, T-statistic, F-statistics , P-values etc.

Min-Max Scaling (Normalized Scaling) :-

- Minimum and Maximum value of features used for scaling. Scale value is between (0 and 1) and (-1 , 1)
- It is highly effected by the outliers
- MinMaxScaler from sklearn is used to normalize .

Standardized Scaling :-

- Mean and Standard deviation is used for scaling. It is used when you want to ensure zero mean and 1 std unit deviation.
- Less effected by the outliers
- StandardScaler from sklearn is used to normalize.

Q 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans. VIF (Variance Inflation Factor) : It's measure how much the behavior of independent variables is influenced by its interaction/correlation with the other independent variables. Its provide a measure of Multicollinearity among the independent variables in MLR.

VIF is infinite define the perfect correlation between two independent variables. Means its represent the Multicollinearity . both independent variable effect the same way so we can drop one of them it will not affect the result.

VIF infinity means one variable is expressing the same exactly by a linear combination of other variable.

Q 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans. Q-Q plot is graphical plotting of the quantiles of two distributions with respect to each other. Whenever we are interpreting a Q-Q plot, we shall concentrate on the y=x line.
For example , Median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plot is to find out if two sets of data come from the same distribution. Plotting the first data set's quantiles along the x-axis and plotting the second data set's quantiles along the y-axis is how the plot is constructed.

Q-Q plots can be used to determine skewness as well. If the left side of the plot deviating from the line called left-skewed and when the right side of the plot deviates, called right-skewed.

We will use Statsmodels.api for Q-Q plot.

If the dataset we are comparing are of the same type of distribution type, we would get a roughly straight line. Below is an example of normal distribution.