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)**

   Ans:

- **season** - Demand is highest in fall, though the medians for both fall and summer are remarkably close. On the other hand, demand falls drastically in spring
- **yr** - Data indicates that demand increased in 2019 as compared to 2018
- **mnth** - Demand appears to be highest between May to October. This also matches the season data since Summer starts in June and Fall starts in September.
- **holiday** - Data indicates a higher demand when the day is not a holiday though 75% quartile for both holiday and non-holidays is the same
- **weekday**- This plot shows almost equal demand from median on any day of the week though the 75% quartile is slightly higher on thursdays, Fridays and Sundays
- **workingday** - Data appears to indicate that there is not much difference in median whether day is a working day or not.
- **weathersit** - weathersit plot appears to indicate user preference of using bikes when the weather is clear or in case of mist with little demand when there is light snow and absolutely no demand with inclement/very bad weather conditions like heavy rains, thunderstorms and snow. This also corresponds to the season and mnth plots which are typically indicative of these weather situations.

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

Ans: When creating dummy variables, the data in the column is split into multiple columns with each column being the individual data values of the column and marked as '1' if the value matches and '0' if the value does not.

For example: Splitting column 'Grade' with grades A, B and C will create 3 columns Grade_A, Grade_B and Grade_C with values as 1 if the row value matches for that particular and 0 if it does not match.

In this case, it can be noted that when there are n values, n variables get created. However, we can explain the values of n variables with just n-1 dummy variables.

For example: In the previous grades example, if Grade_B value = 1, Grade = B; if Grade_C value = 1, Grade = C and if Grade_B and Grade_C are 0, then Grade = A.

This would make having all n columns redundant.

This has been taken care of with 'drop_first =True' in the pandas library.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

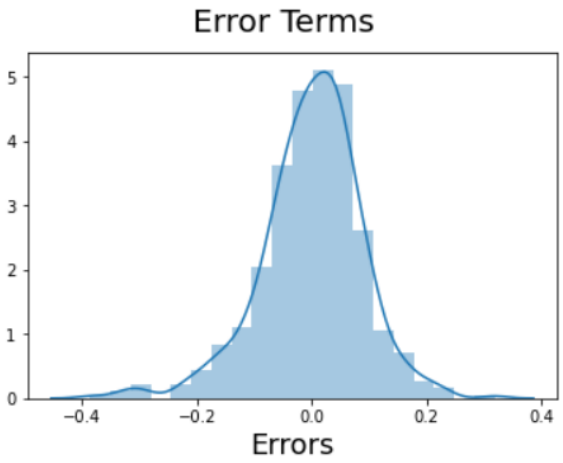
Ans: 'registered' displays the highest correlation, since target variable 'cnt' is sum of 'registered' and 'casual' users.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Ans: The following assumptions of linear regression have been tested:

1. Assumptions about the residuals:

Normality assumption and zero mean assumption: It is assumed that the error terms, ε(i), are normally distributed. – This was tested as part of residual analysis by plotting a histogram of the error terms



2. Assumptions about the estimators:

The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data – This is tested using VIF. In most scenarios, VIF <10 is considered as a good estimator of there being no multicollinearity between variables.

Features	VIF	
1	temp	2.06
2	hum	1.88
7	mnth_Jan	1.55
12	weathersit_Mist	1.55
5	season_winter	1.53
8	mnth_Jul	1.43
4	season_summer	1.41
11	weathersit_LightSnow	1.24
3	windspeed	1.19
9	mnth_Sep	1.19
6	yr_1	1.03
10	holiday_1	1.02

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: Looking at the coeffs, the three features contributing significantly to demand are:

Temp (coef = 0.5717), weathersit_LightSnow i.e. in case of 'Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds' (-0.2362) and windspeed (coef = -0.1945)

This means that demand is positively correlated to the temperature, as temperature rises, so does the demand. Similarly, there is a negative correlation between presence of 'Light Snow and Light Rains' weather condition and demand as well as between 'windspeed' and demand.

Year 2019 did see a rise in demand with coef= 0.2289, however, the year number itself may not figure out in deciding the demand.

General Subjective Questions

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

Ans: Linear regression is a form of machine learning where we train the model to predict behaviour of the data based on certain variables.

Linear regression assumes a linear correlation between the variables on X- axis and Y-axis.

For example, checking relation between sales promotion and increase in number of customers based on previous data related to sales promotion and customer numbers. The idea here is that since we have the historical data, we create a model which will help predict the numbers for the future by learning the behaviour and patters from the historical data.

Linear regression is used to predict effect of independent variable(variables) X on the quantitative response y.

A linear equation can be described as:

y = B0 +B1*X

where y is the dependent target variable

B0 is the slope of the line

B1 is y-intercept of the line

X is the independent variable(or variables)

Linear regression algorithm follows these steps:

1- Splitting the known historical data as train and test.
2- Splitting the train data as the independent (or predictor) variable(s) X and the dependent (or target) variable y.
3- Scaling the data i.e. ensuring all the predictor variables are on the same scale.
4- Creating a model from the train data post fine tuning to find the best model based on factors like R-squared (which measures how well the variance is explained by the fitted line), p-values (to check which coefficients i.e. y-intercept values are significant i.e. value close to zero), Prob(F-statistic) which explains helps measure the fit of the line, etc.
5- The model thus created is used against the test data (similar splitting of X and y and scaling is done with test data as well) to predict the values of y.
6- We then cross check to ensure that the predicted values and actual values are similar and the train and test values of the statistics are similar.
7- This model can now be used to predict values in actual data for which values of y need to be predicted.

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

Ans: Anscombe's quartet was constructed by statistician Francis Anscombe in 1973 to demonstrate the following:

- The importance of graphing data before analyzing it
- Effect of outliers and other influential observations on statistical properties.

The aim of this demonstration was to counter the belief among statisticians that 'numerical calculations are exact, but graphs are rough.'

Anscombe’s quartet comprises of four data sets each with 11 (x,y) points; that have nearly identical simple descriptive statistics yet very different distributions and appear very different when graphed.

For all four datasets:

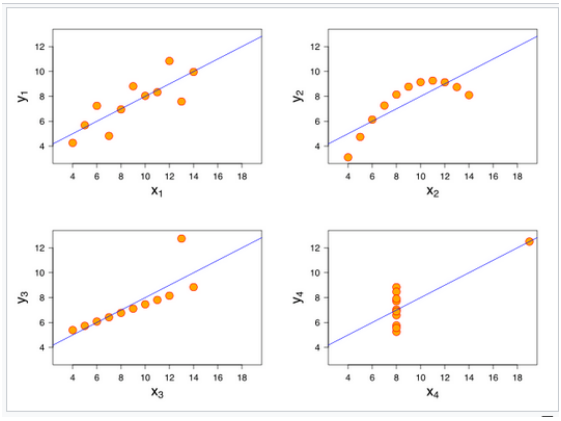
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	y = 3.00 + 0.500x	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R-squared	0.67	to 2 decimal places

The datasets are as follows. The x values are the same for the first three datasets.

Anscombe's quartet data

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

Anscombe’s quartet graphs:

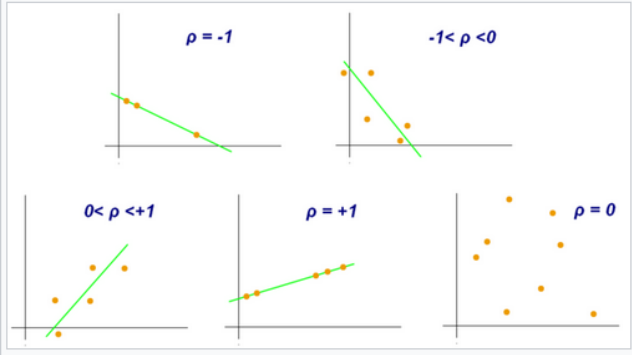


The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.
The quartet is still relevant in that it is used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

Ans: Pearson's R, also referred as the Pearson correlation coefficient (PCC), the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.



Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

For a population:

Pearson's correlation coefficient(ρ (rho)), when applied to a population may be referred to as the population (X,Y), the formula for ρ is:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Eq.1)}$$

where:

cov is the covariance
σ_X is the standard deviation of X
σ_Y is the standard deviation of Y

For a sample:

Pearson's correlation coefficient, when applied to a sample, may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{(Eq.3)}$$

where:

n is sample size

x_i, y_i are the individual sample points indexed with i

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: Scaling referred to as Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Rescaling is required to ensure variables have a comparable scale.

If we do not have comparable scales, then some of the coefficients obtained will be very large or very small compared to others making it very difficult to compare during model evaluation. Without scaling, if values are very high coefficients are smaller and vice versa. In this case if coefficients are much larger, can't say they are more important.

2 commonly used techniques to perform Feature Scaling:

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1. Takes care of outliers.
  $$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$
- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1. Doesn't compress data between particular range. Especially useful in case of extreme outliers.
  $$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

Advantages of scaling:

1. Interpretability
2. Gradient Descent converges faster

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans: Variance Inflation Factor(VIF) is defined as 1/(1-R-squared).

VIF is used to test for multicollinearity that is collinearity between the independent variables.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)

Ans: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus, the line is a parametric curve with the parameter which is the number of the interval for the quantile.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.