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)

   Answer:

   The following conclusions are based on the results of the analysis with 11 features, which all together have significant contributions to the model. In other words, the following factors might booster the sales of the bike sharing.

   1. When weather conditions are: no light rain/snow, no mist/cloud, or having a lower windspeed, the sales are likely much better.
   2. During the seasons in summer, fall, or winter, the sales are better except in spring.
   3. In September and October, the sale trends are doing better.
   4. During working days and Saturdays, the sales tend to be good.
   5. The sales in year 2019 are better than the ones in 2018. Following the sale trend, the sales in following years are likely doing better.

   The following are 11 features that have significant contributions to the models:
   - wsit3_light_rain_snow
   - windspeed
   - wsit2_mist_cloud
   - workingday
   - W6_saturday
   - mnth09_sep
   - mnth10_oct
   - season4_winter
   - yr
   - season2_summer
   - season3_fall

   - According to the results, the equation of our best fitted line is:
   CNT = (-0.317)* light_rain_snow + (-0.175)*windspeed + (-0.093)*mist_cloud + 0.058*workingday + 0.066*Saturday + 0.082*sep + 0.098*oct + 0.190*winter + 0.248*yr + 0.255*summer + 0.294*fall

   - Overall, we have a decent model with the value of R^2 at 77.8% in training data and at 77.1% in test data.

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

Answer: Dummy Variable Trap is a problem that occurs when dummy variables are created in regression analysis. However, the trap can be avoided if one of the dummy variables is dropped or a different coding method is used. By doing so, it will prevent perfect multicollinearity. In other words, if we do not use drop_first = True, all number of dummy variables will be kept. Those dummy variables will be themselves correlated, which cause "multicollinearity". It will turn out to lead to Dummy Variable Trap. Therefore, it's important to drop one of the dummy variables once they are created.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable? (1 mark)
   Answer: The "atemp" numerical variable has the highest correlation with the target variable "CNT".

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Answer:
   The following should be examined:
   1. Normality of error terms: It is important to review the graphical residual distributions to make sure if it is normally distributed with a mean of 0.
   2. Homoscedasticity: A scatterplot between dependent and independent variable can be examined to see if the variance between the predicted and observed values will be a constant for any independent variable.
   3. P-value of F test: The variable with p-value of variable greater than .05 will be removed as it has no significant contribution for the model prediction.
   4. Multi-collinearity: The variable with a value of VIF greater than 5 will be removed.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Answer:
   Based on the absolute value of the coefficients from this model, the top three features with significant contributions for the model are:
   1. wsit3_light_rain_snow
   2. season3_fall
   3. season2_summer

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.778
Model:	OLS	Adj. R-squared:	0.773
Method:	Least Squares	F-statistic:	159.0
Date:	Mon, 19 Jun 2023	Prob (F-statistic):	4.56e-155
Time:	19:12:51	Log-Likelihood:	422.73
No. Observations:	510	AIC:	-821.5
Df Residuals:	498	BIC:	-770.6
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2337	0.019	12.515	0.000	0.197	0.270
yr	0.2476	0.010	26.039	0.000	0.229	0.266
workingday	0.0583	0.013	4.493	0.000	0.033	0.084
windspeed	-0.1751	0.029	-6.008	0.000	-0.232	-0.118
wsit2_mist_cloud	-0.0934	0.010	-9.204	0.000	-0.113	-0.073
wsit3_light_rain_snow	-0.3173	0.029	-10.971	0.000	-0.374	-0.260
season2_summer	0.2552	0.014	18.785	0.000	0.229	0.282
season3_fall	0.2941	0.014	20.347	0.000	0.266	0.323
season4_winter	0.1897	0.016	12.218	0.000	0.159	0.220
W6_saturday	0.0661	0.017	3.951	0.000	0.033	0.099
mnth09_sep	0.0820	0.019	4.342	0.000	0.045	0.119
mnth10_oct	0.0982	0.020	4.864	0.000	0.059	0.138

General Subjective Questions

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

Answer:

Linear regression is one of the common machine learning algorithms under supersized learning. The most common types of linear regression are simple linear regression and multiply linear regression.

Simple linear regression model is commonly used to identify the relationship between a dependent variable (y) and an independent variable (x) with a straight line to mode the linear association, as known as “linear regression”. In the regression model, the independent variable is also modeled as predictor, while the dependent ones are modeled as output variables.

Mathematically, the relationship can be expressed in the following equation:

$Y = c + m*X$

- Y is the dependent variable as the target one being predicted.
- X is the independent variable being used to predict.
- m is the slope of regression line, which represents the relative changes between X and Y when one value changes, another will follow too.
- C is a constant, knows as Y-intercept. When X=0, Y = C.

The linear relationship can be positive or negative between independent and

dependent variables. Positive linear relationship can be defined when one variable increases, another will increase and vice versa. Negative linear relationship can be defined when one increases, another decreases.

Assumptions

Following assumptions about the dataset are made when linear regression model is assumed:

1. Minor or no multi-collinearity: Multi-collinearity occurs when there are high intercorrelations among two or more independent variables in a multiple regression model.
2. Linear relationship between response and feature variables.
3. Normality of error terms: in linear regression, error terms are always expected as in real life independent variables are never perfect predictors of the dependent ones.
4. Homoscedasticity: Residual variances should be morally distributed.

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

Answer: Anscombe's quartet developed by a statistician by Francis Anscombe presented four very different distributions from four data sets, which have almost identical statistical properties. The differences can be easily detected by plotting the two variables on four different figures. Anscombe's quartet suggests the data features must be plotted with visual presentations to show the data distribution that can help identify various anomalies occur in the data like outliers, diversity of the data, linear separability of the data, etc.).

As shown below in Figure 1 presented by Graeme L. Hickey and others (2015), in the study, the authors indicated that each dataset comprises 11 data points (orange points) with nearly identical statistical properties, including means, sample variances, Pearson's sample correlation (denoted as r=0.82 in the figure), and linear regression line (blue lines).

Figure 1: Anscombe's quartet with graphical presentations.

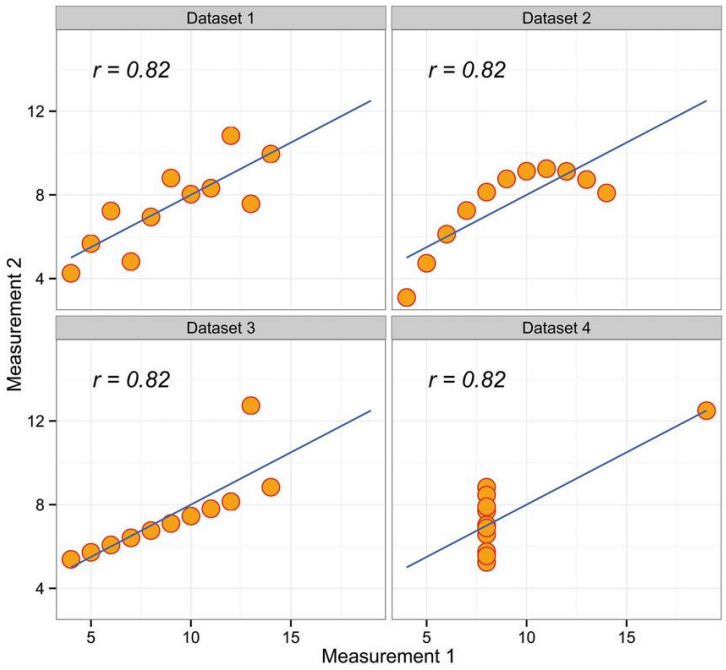

Figure 1 is adapted from the source:
Graeme L. Hickey and others, on behalf of the EJCTS and ICVTS Editorial Committees, Statistical and data reporting guidelines for the *European Journal of Cardio-Thoracic Surgery* and the *Interactive CardioVascular and Thoracic Surgery*, *European Journal of Cardio-Thoracic Surgery*, Volume 48, Issue 2, August 2015, Pages 180–193, https://doi.org/10.1093/ejcts/ezv168
https://www.researchgate.net/figure/Scatterplots-of-four-different-datasets-known-as-Anscombes-quartet-99-Each-dataset_fig1_276360680

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

Answer:

Pearson correlation coefficient is the measure to express how closely two variables are related in a linear association. It ranges from -1 to 1, where -1 indicates a perfect negative association. It means that if one variable increases, another will decrease. A value of 0 means no relationship. A value of 1 indicates a perfect positive relationship. It means that if one variable increases, another will increase too.
The coefficient is calculated by dividing the sum of the products of the deviations of each variable from their mean by the product of their standard deviations.
Its formula can be expressed as:

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

- r = Coefficient of correlation
- xbar = Mean of x-variable
- ybar = Mean of y-variable.
- x_i y_i = Samples of variable x, y

(The formula is adapted from https://www.simplilearn.com/tutorials/statistics-tutorial/pearson-correlation-coefficient-in-statistics)

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

Answer:

Feature scaling is a strategy to rescale the ranges of independent variables through data normalization or standardization procedure. It is typically conducted during the data preprocessing stage to minimize the range differences among varying units of independent variables. 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, irrespective of the units of the values. Doing so to ensure that all features contribute equally to the model. The data could be transformed to a more consistent scale that help build accurate and effective machine learning models.

Normalization is commonly used, when you believe that the data distribution does not follow the Gaussian distribution. It's especially helpful for the data that seems not normally distributed or an unknow distribution like using K-Nearest Neighbors and Neural Networks. On the other hand, standardization is still helpful for whether the data follow a Gaussian distribution or not. Also, unlike the normalization method, standardization is preferred sometimes, because it doesn't set a bounding range between 0 and 1. By doing so, the outliers still can be observed in your data, because they will not be affected by standardization.

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

(3 marks)

Answer:

The formula of VIF= $$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where:

R_i^2 =Unadjusted coefficient of determination for regressing the i^{th} independent variable on the remaining ones (Source adapted from: https://www.thetechplatform.com/post/variance-inflation-factor-vif)

If R^2 is 1, the VIF will be infinite. It also means that there is a perfect correlation between two features. To solve this problem, there is a need to drop one of the variables that causes the perfect multicollinearity.

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

(3 marks)

Answer: The Q-Q plot is the visual presentation using a scatter plot to show two sets of quantiles against one another. It aims to check if two sets of data came are normally distributed. When they are normally distributed, the plotted data points appear to be

closely along the diagonal line. The data points significantly deviating from the straight line are indicative of outliners.