Assignment-based Subjective Questions

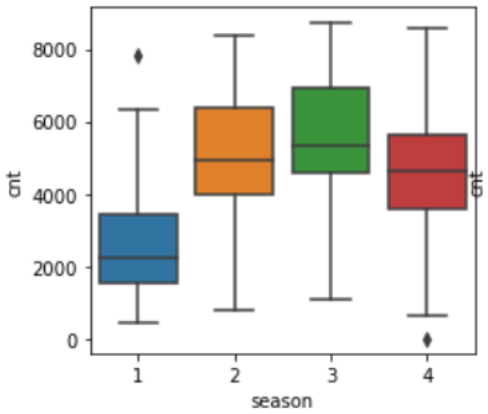
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Categorical variables: Below are the categorical variables identified from the Boom Bikes dataset

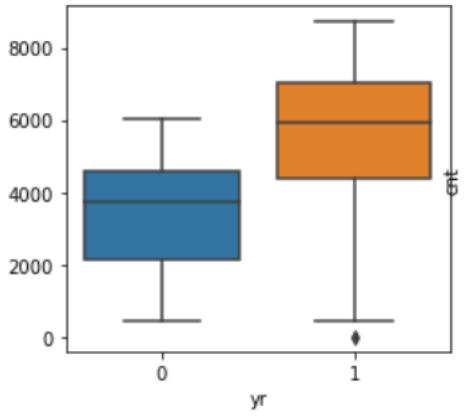
```
cat_var = ['season','yr','mnth','holiday','weekday','workingday','weathersit']
cat_var
```

```
['season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', 'weathersit']
```

Season : Rental bikes median cnt increased during summer and fall seasons with 4941 and 5353 respectively and remained lowest during spring season with median cnt 2222. With this we can infer that rentals cnt increase during summer and fall seasons.



yr: During the base year 2018 the rental bike median cnt is 3740 and it has increased to 5936 in the year 2019

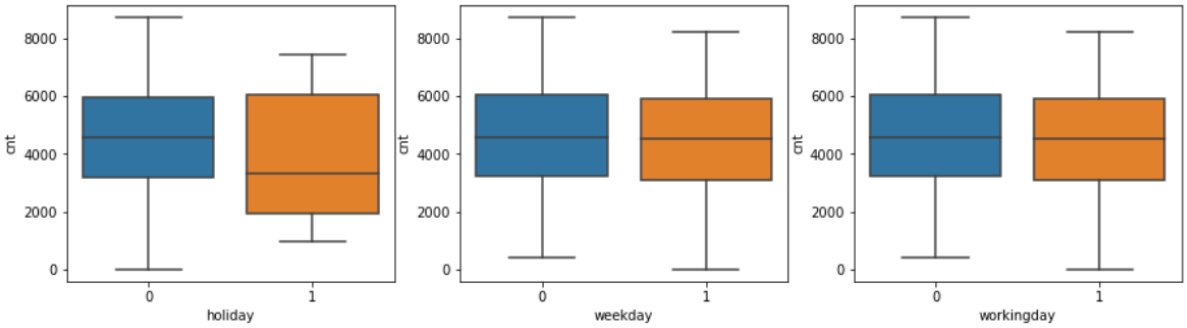


holiday, working day, weekday :

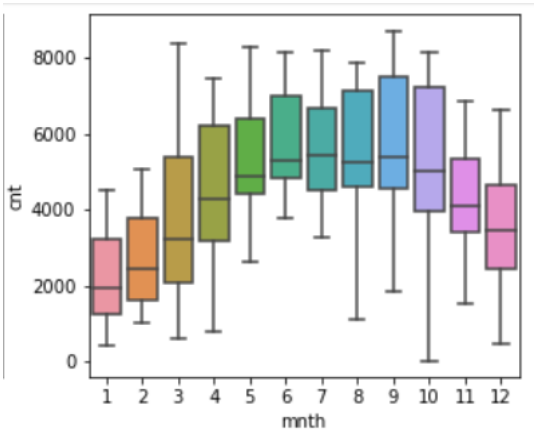
- During the holidays the rental bikes median cnt is higher as compared to non holidays.
- The working days rental bikes median cnt is equal to not working days rental bikes median cnt

- Weekdays median count is equal to weekends weekdays

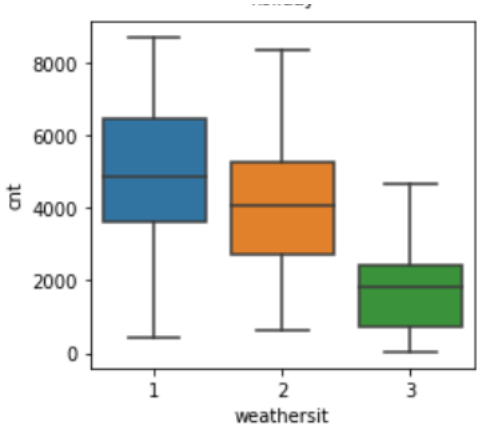
Inference: During the holidays and weekends the rental count increases. Holidays and weekends compensate or contribute equally to working days or weekdays.



mnth: median rental bikes count increases to peak during the months of may, June, July august, sept and then decreases from oct onwards.



weathersit: As the weathersit moves from clear or little clouds to heavy rain with icepallets or thunderstorm the rental bikes cnt decreases which you can observe in the below figure. From this we can infer that bad weather has an negative impact on the number of rental bikes and good weather conditions bring more bike rentals.

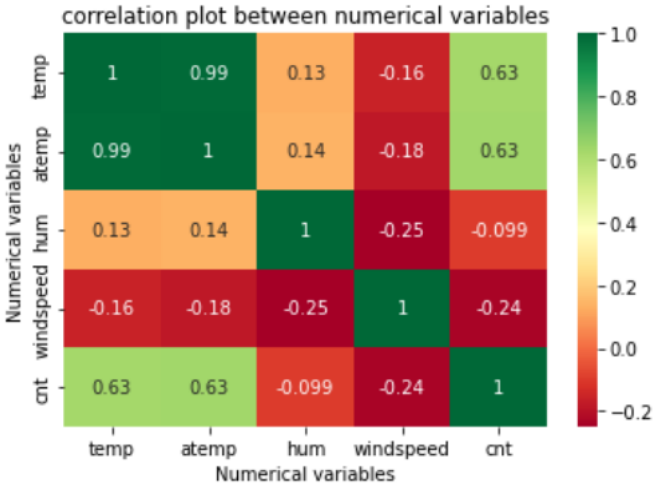


2. Why is it important to use drop_first=True during dummy variable creation?

During dummy variable creation we use drop_first = True because first column is reduntant and the rest of the remaining columns captures the effect of first column. But in our Boom bikes case study since the company wants to know which features or columns have impact on the target variable and for interpretation purpose we do not want to use drop_first = True but try to study the collinearity effect and variance inflation factors for those first dummy columns and then eliminate.

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

Temp and atemp variables have the highest correlation of 0.63 with the target variable cnt and the relation is positively correlated.



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

1. Variance inflation factor - Minimised Multicollinearity
2. Linear relationship between predictor variables and target variable
3. Error terms (Homoscedascity check, normal distribution,linear Q-Q plot)
4. No correlation between Independent variable(temp , windspeed) and residual.

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

The Top 3 features are temp, yr_1, weathersit_3 with coefficients 0.5173, 0.2326, -0.2819 respectively are the top 3 significant features.

General Subjective Questions

Explain the linear regression algorithm in detail.

Linear regression is about finding a best fit line which captures the statistical linear relationship between the target variable and one or more predictor variables. The best fit line is the one which minimises the prediction error. The linear regression explains the change in target variable with the change in predictor variable.

1. The equation for the linear regression algorithm is of the form

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \ldots\ldots\ldots \beta nXn$$

X1 ,X2, X3 ,X4,….Xn are predictor variables ; (can be one or many predictor variables)

Y is target variable or dependent variable,

β0, β1 β2 β3 β4…. Bn are the coefficients.

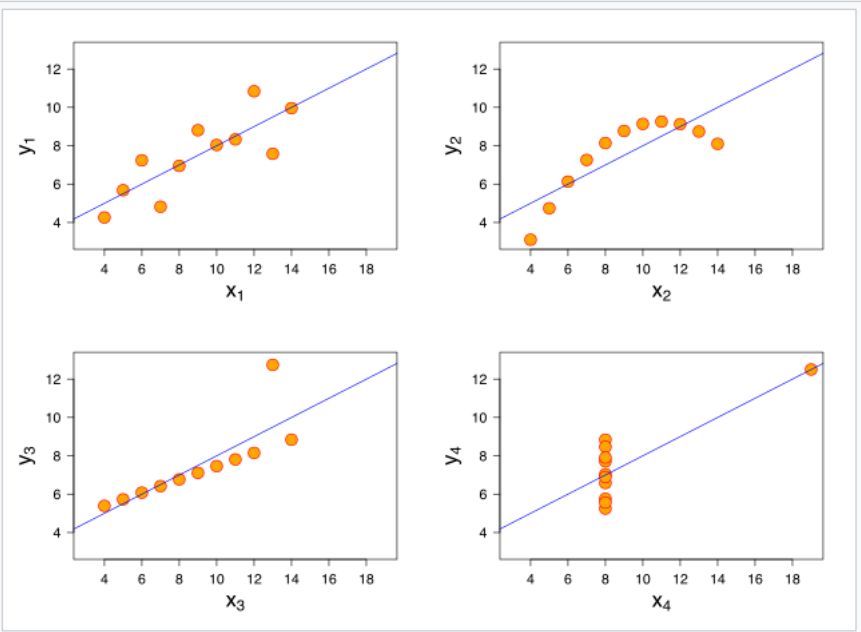
2. Linear regression shows the correlation between variables but not causation.
3. In the industry linear regression is mainly used for forecasting and prediction. In the prediction the focus is on the driver variables and measuring the impact on the target variable, where as in the forecasting the importance is given to achieving the accuracy in final forecast.
4. The following are the assumptions for algorithm to be linear regression algorithm:
    - Linear relationship between the target variable and the predictor variables.
    - Normal distribution of error terms
    - Independence of error terms to the predictor variables
    - Constant variance of error terms against the target variable
5. Using Ordinary least square method we can build the linear regression model and To determine the significance of coefficients of predictor variables we use p value or t-stat statistics.
6. Residual analysis is done to evaluate the linear regression algorithm.
7. Variance inflation factor can be calculated while doing linear regression to figure out how well the independent variable is able to explain other independent variables combined.

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

The power of visualizations is established by Anscombe to counter the common belief in 1973 that graphs are rough and statistics are correct to interpret the data. To prove this he has taken four data sets such that the descriptive statistics like mean, variance, correlation coefficient, best fit line are exactly same but when plotted as graphs all datasets look different. The outliers, curvatures, variance can be seen visually but descriptive statistics shows all the datasets are similar.

Anscombe's quartet

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



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	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^2	0.67	to 2 decimal places

Source reference : https://en.wikipedia.org/wiki/Anscombe%27s_quartet

References :

Anscombe F. Graphs in statistical analysis. The American Statistician. 1973 Feb;27(1):17–21. Available from: https://www.jstor.org/stable/2682899?seq=1

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

In statistics, Pearson's R is pearson correlation coefficient , is a measure of the strength of the linear relationship between two variables. Pearson's R measure ranges between -1 to +1. An R of -1 indicates a perfectly linear relationship with negative slope, an R of +1 indicates a perfectly linear relationship with a positive slope, an R of zero indicates no linear relationship.

It is ratio between the covariance of the two variables and the product of their standard deviations.

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

Scaling is an approach to bring the dataset variables values onto the same scale or fixed range.

- This helps in improving the performance of some of the machine learning algorithms which uses gradient descent mechanism to reach the global minima, which is essentially a method to find the model which minimizes the errors.
- It also helps to rescale the variables so that they have comparable scales. If some variables have large coefficients and some variables have small coefficients during the model building process it is difficult to evaluate the model.

Normalized Scaling	**Standardised Scaling**
Values rescaled between 0 and 1	values rescaled to mean = 0 and std.dev to 1
Calculated as X_new = (X - X_min)/(X_max - X_min)	Calculated as X_new = (X - mean)/Std
outliers does not impact the scaling	not reliable scaler when outliers are present
Good to use when we do not know the distribution of the dataset	Good to use when dataset follows a normal distribution or gaussian distribution

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

VIF indicates how well the independent variable is correlated to other independent variables combined.

- VIF of infinity implies perfectly correlated as (1/1-R^2) tends to infinity,
- which indicates high collinearity as R^2 tends to 1 and VIF becomes 1/0.
- R^2 equal to 1 implies a perfectly correlated independent variable.

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

Q -Q plot is a plot of sample distribution against the theoretical distribution. This plot helps us to determine if the dataset follows any particular distribution like normal distribution, uniform distribution, exponential distribution.

Machine learning models like linear regression perform better when the features follow normal distribution.

Q-Q plot are useful in linear regression and helps us to determine :

1. If two datasets are of the same distribution
2. For a linear regression we assume that Error term follows a normal distribution when the model is linear, so we can use this Q-Q plot to check this assumption
3. Skewness of the data is determined