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?

Final list of variables in decreasing order of size of impact(positive or negative)

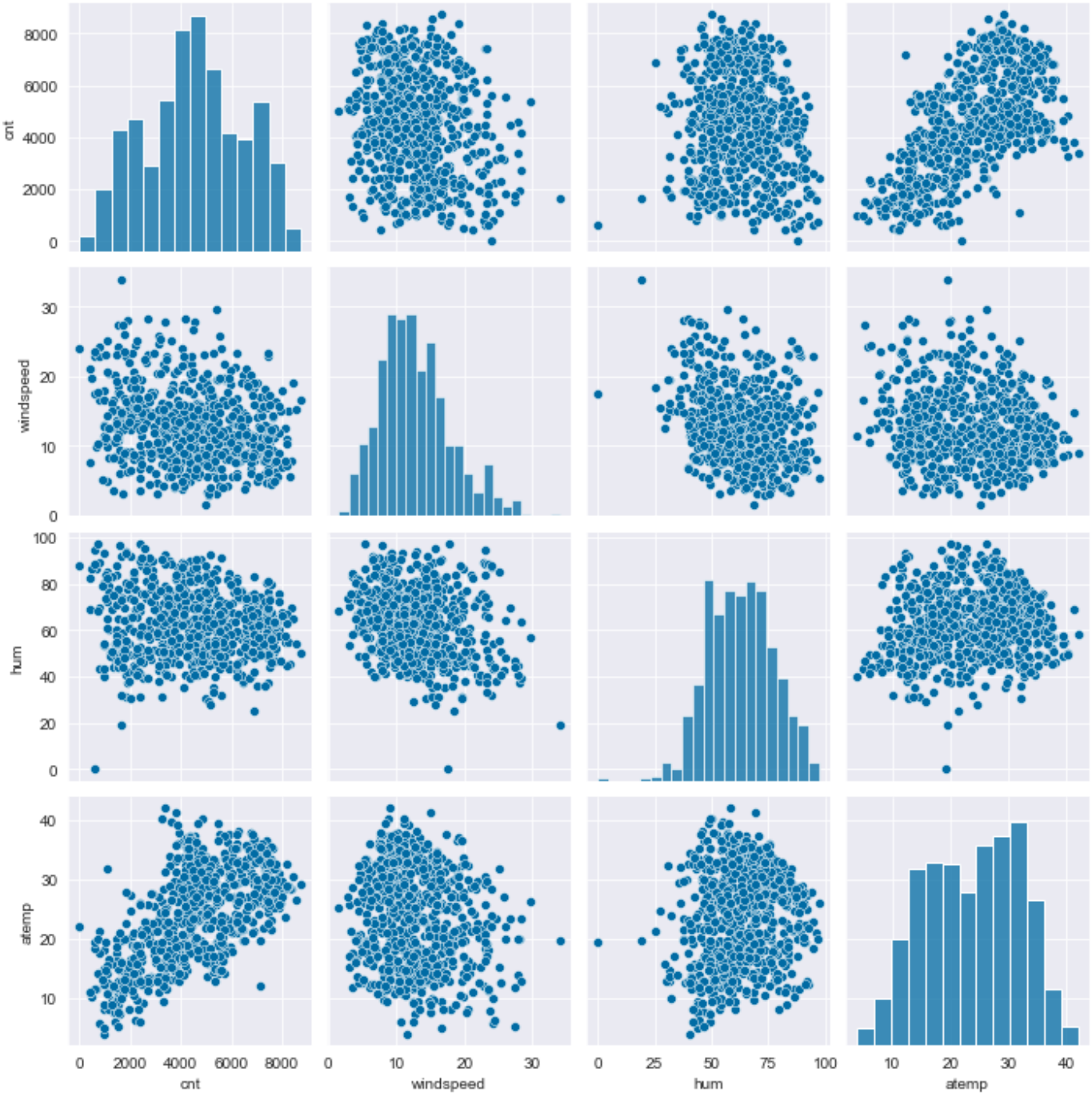
- **weathersit_3[extreme weather conditions for biking]:** -1.330571 (Negative effect)
- **yr[Year 2019]:** 1.051845 (Positive effect)
- **season_4[Winter]:** 0.724146 (Positive effect)
- **mnth_9[Sep]:** 0.538348 (Positive effect)
- **season_2[Summer]:** 0.525488 (Positive effect)
- **atemp[Feels like temperature]:** 0.412762 (Positive effect)
- **holiday[Holiday flag]:** -0.385588 (Negative effect)
- **season_3[ZFall]:** 0.377814 (Positive effect)
- **weathersit_2[Cloudy weather]:** -0.362133 (Negative effect)
- **mnth_8[Aug]:** 0.277839 (Positive effect)
- **mnth_5[May]:** 0.209645 (Positive effect)
- **mnth_6[Jun]:** 0.195376 (Positive effect)
- **mnth_10[Oct]:** 0.177703 (Positive effect)
- **mnth_3[Mar]:** 0.170554 (Positive effect)

Overall we can see that broadly, the factors that have an effect on bike sharing demand are weather & climate conditions, along with some seasonality factors like holiday and the year 2019. All of these factors are what conditions are best suited for customers to ride bikes outside.

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

When we use 'get_dummies' on a categorical variable, by default it will create one new binary 1/0 variable for every level in the categorical variable. Let's say a categorical variable has 'm' levels. By default get_dummies creates 'm' binary variables. It is important to note that , absence(0) of a level in 'm-1' binary columns implies the presence(1) in the 'm'th binary variable. This means that if we have 'm' binary variables, there is linear dependence amongst the variables. Hence we use the 'drop_first=True' condition so that the first level is dropped and we avoid linear dependence amongst the independent variables. In simple terms, 'drop_first=True' is used to avoid multicollinearity.

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



The pairplot is shown above. The variable with the strongest correlation with 'cnt' are the temperature variables - 'temp' and 'atemp'. Amongst them temp and atemp have 99+% correlation

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

There are different ways in which we can validate the assumptions of linear regression on a model built on training set:

- pairplot/scatter plots of X & y will tell us if there is linear relationship between the dependent vriable and the independent variable
- we leverage the correlations and vif calculation and iterativley select variables in the model to ensure we are not using multi collinear variables
- Each independent variable must be close to normal distribution. Else we need to do transformations accordingly

- Calculate residuals(actual minus predicted) and do a distribution plot. The residuals must be normally distributed centered around zero
- plot residual with the distribution of the variable. The variance in residuals should be similar across the range of the variable. This can be visually inspected
- For autocorrelation, Durbin-watson test is used

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

The top positive contributing factors are:

- Yr 2019
- Season[winter]
- month[Sep]
- Season[Summer]
- temperature

Also, note that 'extreme weather conditions' have the most negative impact

Final list of variables in decreasing order of size of impact(positive or negative)

- **weathersit_3[extreme weather conditions for biking]:** -1.330571 (Negative effect)
- **yr[Year 2019]:** 1.051845 (Positive effect)
- **season_4[Winter]:** 0.724146 (Positive effect)
- **mnth_9[Sep]:** 0.538348 (Positive effect)
- **season_2[Summer]:** 0.525488 (Positive effect)
- **atemp[Feels like temperature]:** 0.412762 (Positive effect)
- **holiday[Holiday flag]:** -0.385588 (Negative effect)
- **season_3[ZFall]:** 0.377814 (Positive effect)
- **weathersit_2[Cloudy weather]:** -0.362133 (Negative effect)
- **mnth_8[Aug]:** 0.277839 (Positive effect)
- **mnth_5[May]:** 0.209645 (Positive effect)
- **mnth_6[Jun]:** 0.195376 (Positive effect)
- **mnth_10[Oct]:** 0.177703 (Positive effect)
- **mnth_3[Mar]:** 0.170554 (Positive effect)

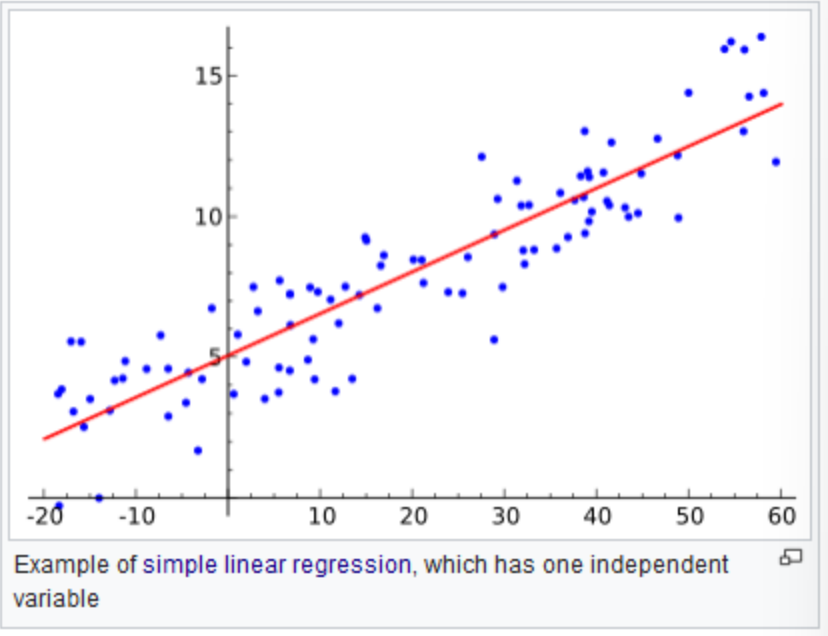
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is one of the most basic methods for regression used to predict numeric dependent variables. In linear regression, each of the independent variables are linearly related with the dependent variable and the coeffcient of the independent variable indicates the size of the effect of the variable. The independent variables can be numeric or binary. Categorical variables need to be converted to binary variables. If there are multiple independent variables, then it is called 'Multiple linear regression' and if its only one variable then it is called 'simple linear regression'.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$$

Loss function in linear regression is a function of the square of error(actual - predicted). Popular approach used are ordinary least squaress and gradient descent to arrive at the parameters of the model.



Example of simple linear regression, which has one independent variable

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a classic example which demonstrates that even if the descriptive statistics are same, the underlying distributions can be vastly different. In the quartet, 4 example distributons of x,y are taken. Each x with a mean of 9,variance of 11 and y with a mean of 7.5, variance of 4.125. Looking att he below, we can understand the signifiacne of visualising the data to understand it better. This also helps us in understanding the influence of outliers and other extreme values.

Datasets:

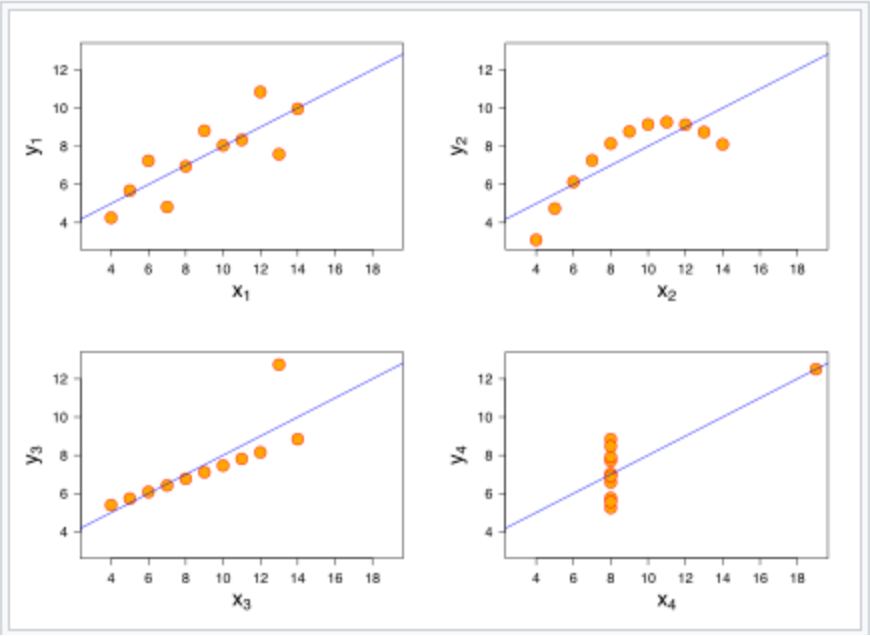
Anscombe's quartet

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

Descriptive stats:

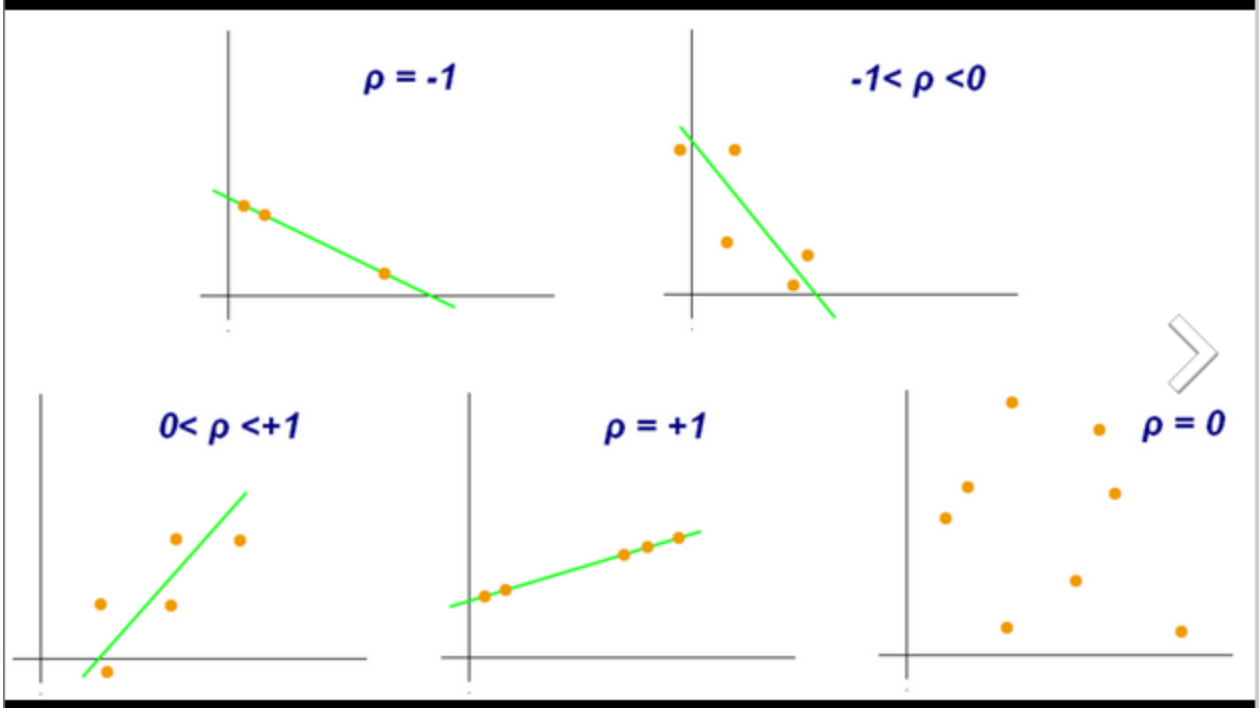
Property	Value
Mean of x	9
Sample variance of x : s_x^2	11
Mean of y	7.50
Sample variance of y : s_y^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression : R^2	0.67

Graphs:



3. What is Pearson's R?

Pearson's R is also called as correlation coefficient. It represenst the strength of linear relationship between the two data points. Numerically, it is calculated as the ratio between covariance of the variables divided by the product of standard deviations. The correlation factor can range btween -1 and 1 with higher the abosulte value of the pearson's R, stronger the relationship. The positive and negative indicates the direction of relationship between the two variables.



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

Numerical variables are at different scales. When we run regression models or use numerical variables in any other models, the scale of the variable dictates the value of the parameter. Also with numeric variables at different scale, the loss function & gradient descent will run slower if numeric variables are at their original scale.

To avoid these problems, we using 'scaling' to bring all numeric variables to have simialr scale. There two methods generally used - Normalisations & Standardisation

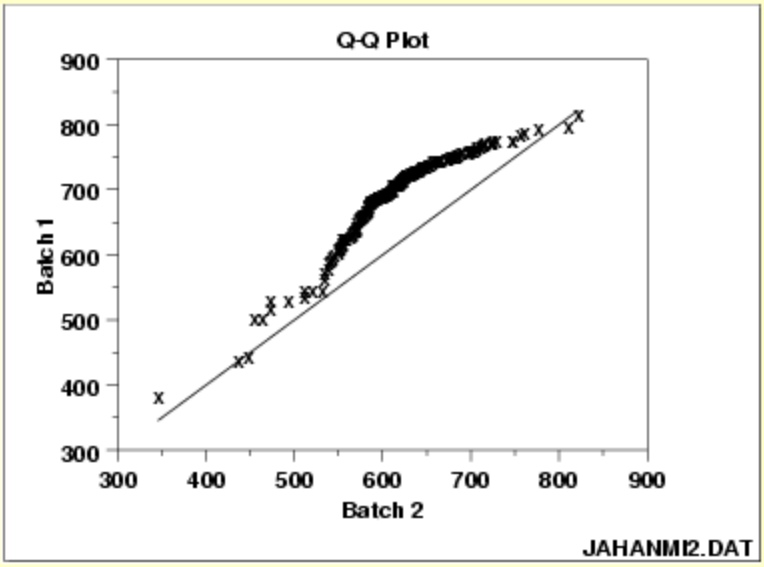
In normlaisation , we scale the variables between 0 and 1 by subtracting 'min' and dividing by range(max-min) in Standardisation, we bring the variable to a mean of zero and standard deviation of 1. this is done by subtracting mean and dividing by standard deviation.

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

The formula for VIF is 1/(1-R2). If the R2 is 1, only in that case VIF is infinity. This happens when there is linear dependence between the features. In other words, if one variable is linearly explained by other variables, then the R2 will be 1 leading VIF becoming infinity.

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

QQ plot is Quantile-Quantile plot. This is a plot of two variables. one on axis we plot quantile values of one variable. On the other axis, we plot the quantile values of another variable. This plot is used to check if two variable distributions come from the same populations. In theory, if they come from the same population, the QQ plot must be closer to the 45 degree reference line. Farther the reference line and the qq plot are, stronger the chance of the samples of coming from different distributions.


Q-Q Plot
900
800
700
600
500
400
300
Batch 1
300
400
500
600
700
800
900
Batch 2
JAHANMI2.DAT