Sample Questions – 1

Q1.)
Generate a random sample of size 1000 from log-Normal with $\mu = 5$ and $\sigma = 2.5$ using `set.seed(10221)`

Calculate the empirical quartiles and compare it with the true quartiles of the distribution and comment.

Q2.)
Load the data set `mtcars`. You wish to predict "mpg"

a.) Build a Linear Regression model with NO explanatory variables
b.) Build a Linear Regression model with "wt" as the only regressor
c.) Build a Linear Regression model with log of "wt" and "disp"
d.) Comment on the 3 models built based on R-sq, Adj-R.sq, AIC.
e.) Draw the fitted line in b.) of red colour on the scatter plot (where each "+" is of orange colour)

Q3.)
Execute the below code and answer the following question:

```
set.seed(2919)
s = rnorm(100,5,2.5)
```

Calculate the empirical mode and compare it with the true model as per the nature of the Normal distribution. How can you reduce that discrepancy between the 2?

Q4.)
Can we conclude if the following sample data comes from Poisson Distribution.
Use Chi-sq GoF test to confirm the same.

X	Frequency
0	739280
1	185309
2	23349
3	1937
4	114
5	1
6	3

Q5.)
Load the `Animals` dataset from `MASS` package.
Understand the 2 variables from the help section by executing the code
`?MASS::Animals`

a.) Comment on the relationship between the 2 variables.
b.) Comment on the log relationship between the 2 variables.
c.) Test whether $\rho > 0.6$ stating your Hypothesis and Conclusion clearly
d.) Test whether the mean Body weight is greater than mean Brain weight.
e.) Which of the variable is more likely to follow Normal Distribution? Comment using QQ-Plot.
f.) You think that the log of body weight follows Normal with mu = 3.77 and sigma = 1.5. Superimpose the empirical distribution and the distribution you thought and comment.

Q6.)
A study of educational levels of voters and their political party affiliations yielded following results:

	Party		
Education	A	B	C
SSC	40	20	10
Graduate	30	35	15
PG	30	45	25

Test whether party affiliation independent of education level?

Q7.)
Import the data: Spam Email.csv
You want to predict whether a given email is a Spam or not.

Every row represents an email and every column represents a word and its corresponding cell represents the no. of times that word has been used in an email.

Build a model to predict whether an email is Spam or not.
Try to create multiple models with different regressors and compare them based on AIC and Null Deviance.