Name: Ishita More

UID: 2019130039

TE Comps

Subject: Data Analytics

ISE-Part(II)

Aim:

Probability distributions and hypothesis testing

Dataset:

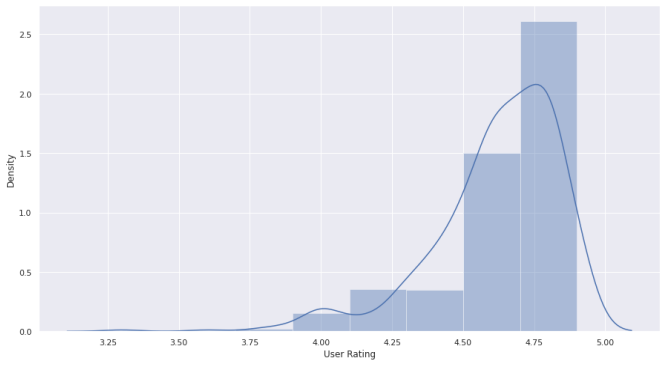
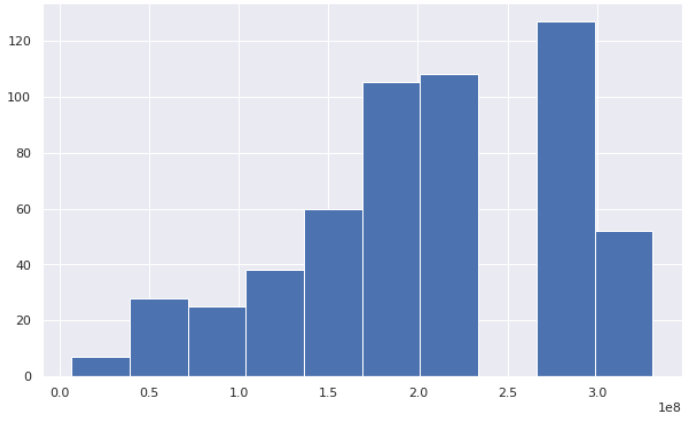
Best Selling Book

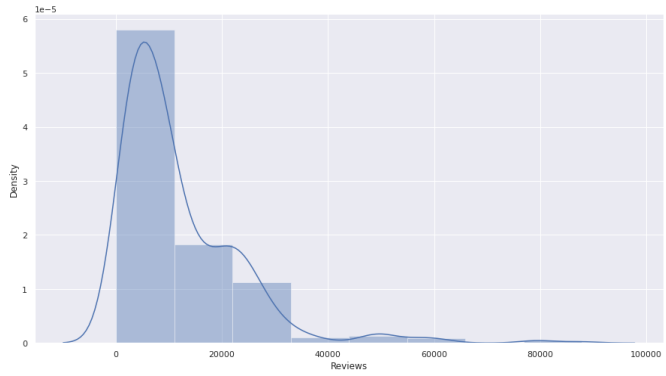
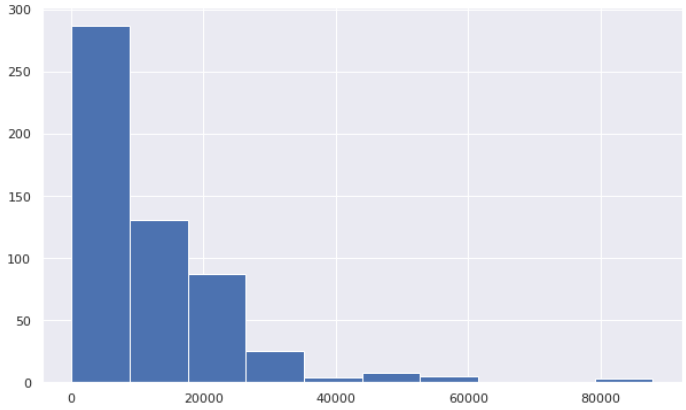
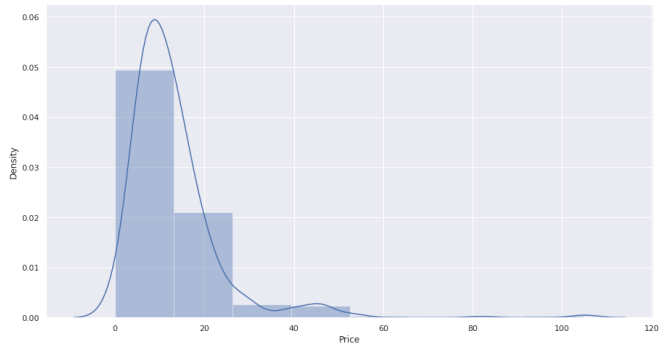
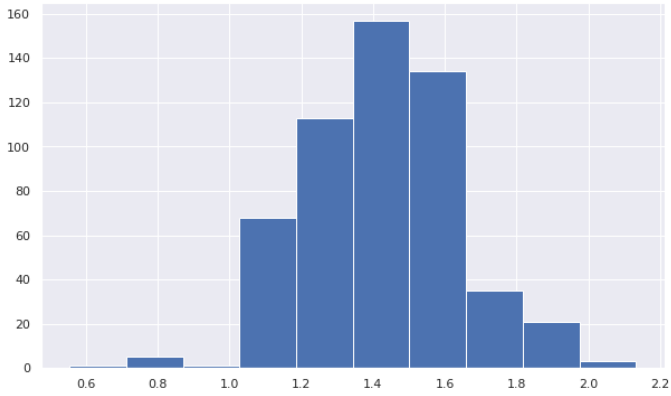
https://www.kaggle.com/sootersaalu/amazon-top-50-bestselling-books-2009-2019?select=bestsellers+with+categories.csv

Dataset on Amazon's Top 50 bestselling books from 2009 to 2019. Contains 550 books, data has been categorized into fiction and non-fiction using Goodreads

Inference:

1.

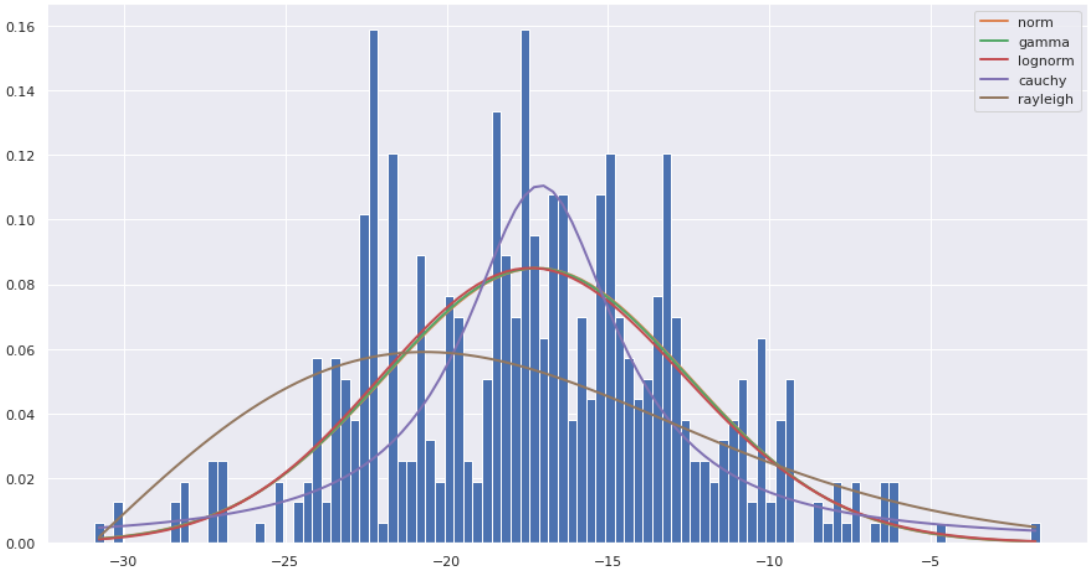
| Dist Plot | Histogram |
|---|---|
| User Rating  |  |

| Dist Plot | Histogram |
| --- | --- |
| Reviews   |  |
| Price   |  |

2.

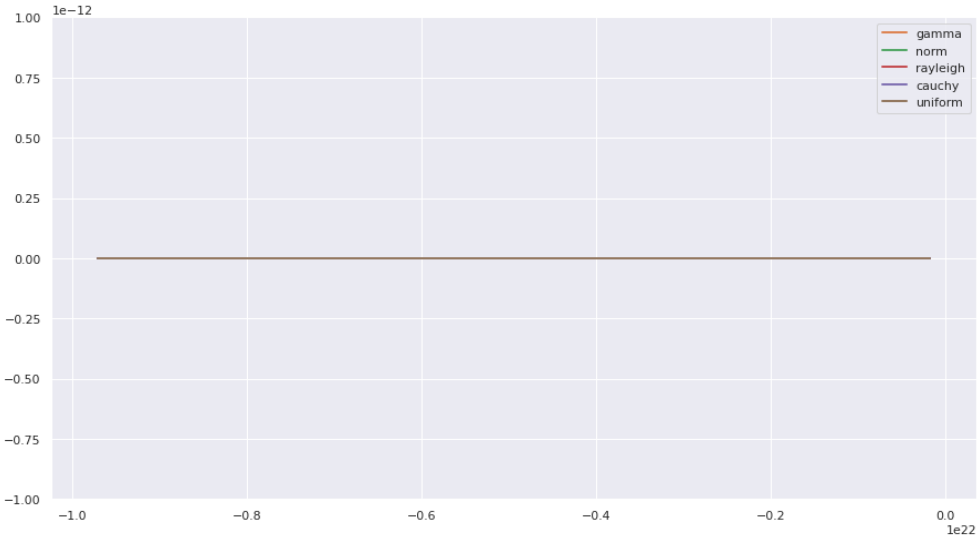
a. For the variable Reviews,

```
tt = data["Reviews"].values
f = Fitter(tt,
        distributions= get_common_distributions())
f.fit()
f.summary()
f.get_best(method = 'sumsquare_error')
```



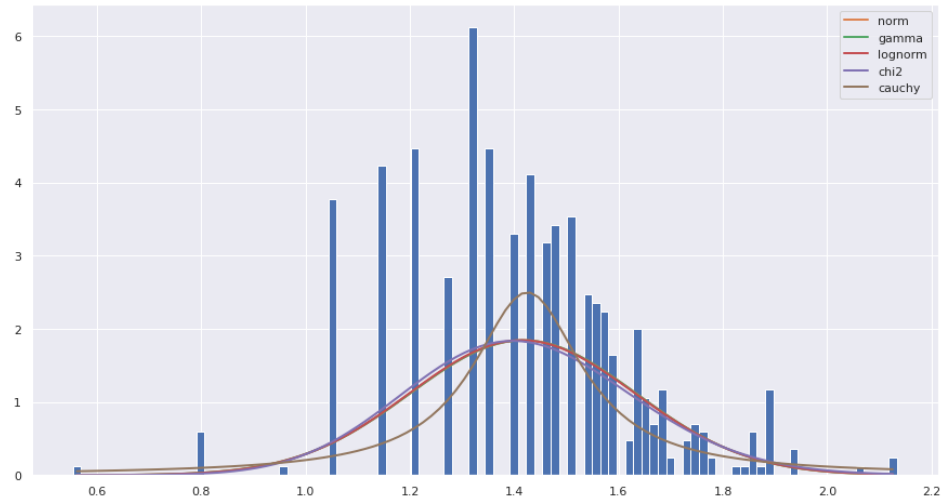

Using the fitter module, we can compute that the best fit for it is Normal Distribution

b. For the variable User Rating,




Using the fitter module, we can compute that the best fit for it is Gamma Distribution

c. For the variable Price,




Using the fitter module, we can compute that the best fit for it is Normal Distribution

3.

A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.

- The first value is the chi-square value of 6961.793 . The chance of such a value or even more extreme, in a sample, if there is no association in the population is 2.30030132739655e-122 (the second value). This is known as the p-value or significance. It is considered 'significant' usually if this value is below 0.05, which in this case it is. This indicates an association between the two variables (one has an impact on the other).
- The third value is the degrees of freedom, which is an indication of the size of the table since it is simply the number of rows - 1, times the number of columns - 1.
- The last array is the so-called expected values. These are the counts to be expected if the two variables had no influence on each other.

H0: (null hypothesis) A variable follows a hypothesized distribution.

H1: (alternative hypothesis) A variable does not follow a hypothesized distribution.

Since the p-value 2.30030132739655e-122 is less than 0.05, we reject the null hypothesis.

Also,

- Since our chi-squared statistic exceeds the critical value, we'd reject the null hypothesis that the two distributions are the same.
- Also, the p_values are less than alpha(0.95), we'd reject the null hypothesis that the two distributions are the same.

In the $\chi 2$ goodness-of-fit test, we conclude that either the distribution specified in H0 is false (when we reject H0) or that we do not have sufficient evidence to show that the distribution specified in H0 is false (when we fail to reject H0).

4.
5.

Using two-sample goodness of fit test i.e. Two-Sample Kolmogorov-Smirnov Test

Ks_2sampResult(statistic=1.0, pvalue=1e-322)

From the output we can see that the test statistic is 1.0 and the corresponding p-value is 1. Since the p-value is less than .05, we reject the null hypothesis. We have sufficient evidence to say that the two sample datasets do not come from the same distribution.

As from above, it is clear that data['User Rating'] is the best fit for gamma distribution, whereas data['Reviews'] is the best fit for normal distribution.

6.

Using data.skew(),

User Rating -1.502125

Reviews 2.421597

Price 3.685057

Year 0.000000

It is concluded that

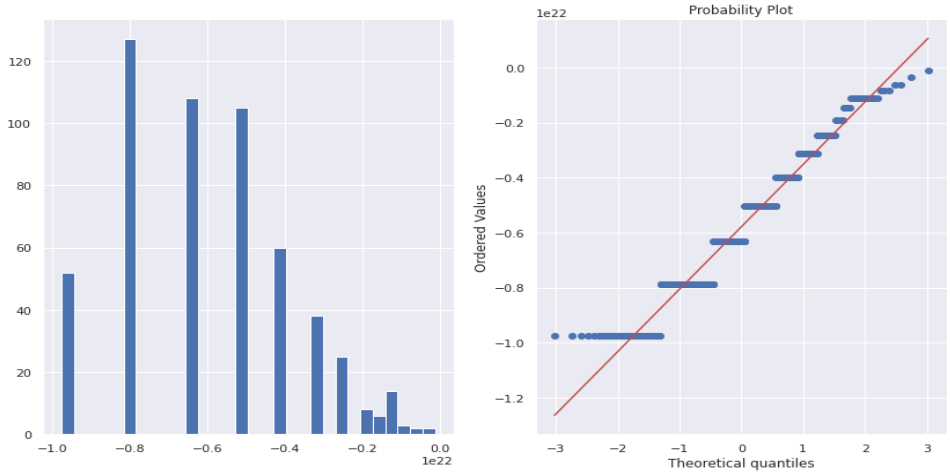
- The variables with skewness > 1 are highly positively skewed.
- The variables with skewness < -1 are highly negatively skewed.
- The variables with 0.5 < skewness < 1 are moderately positively skewed.
- The variables with -0.5 < skewness < -1 are moderately negatively skewed.
- And, the variables with -0.5 < skewness < 0.5 are symmetric i.e normally distributed.

Therefore, it is not normally distributed.

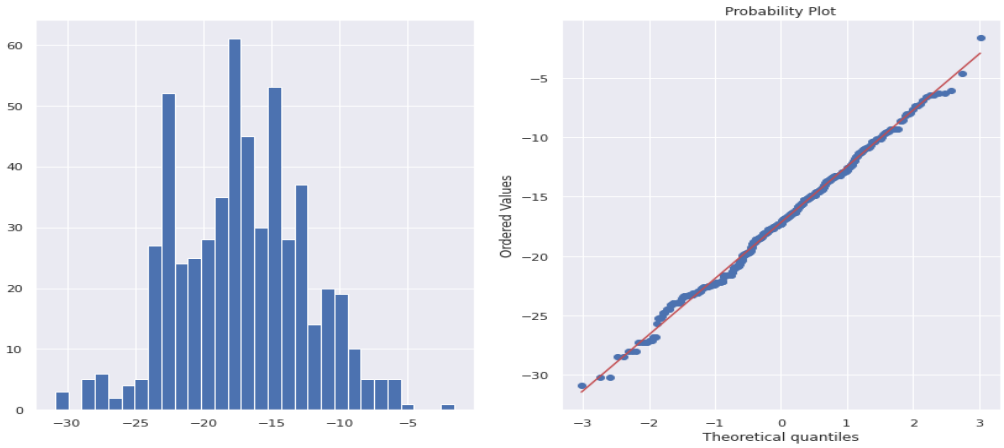
Transformation:

Variable: User Rating

```
data["User Rating"], lambdavalue = stats.boxcox(data['User Rating'])
print("selected value for lambda is " , lambdavalue)
plotvariale(data, "User Rating")
```

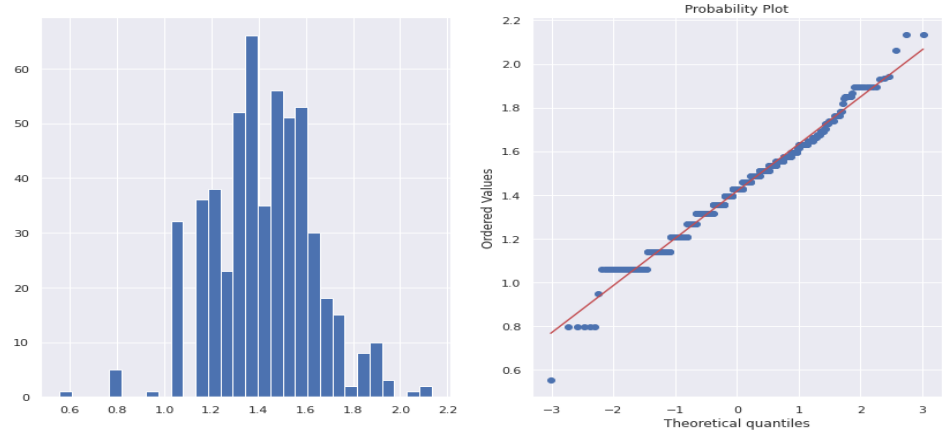



Variable: Reviews




Variable: Price

data['Price'], lambdavalue = stats.yeojohnson(data['Price'])
print("selected value for lambda is " , lambdavalue)
plotvariale(data, 'Price')



data.skew()

```
User Rating    0.121836
Reviews        0.020094
Price          0.000964
Year           0.001073
```

The variables with -0.5 < skewness < 0.5 are symmetric i.e normally distributed.

7.

```
if pval < 0.05:    # alpha value is 0.05 or 5%
   print(" we are rejecting null hypothesis")
  else:
   print("we are accepting null hypothesis")
```

```
User Rating:
-5.75410762494121e+21
p-values:  3.2073785450396196e-232
 we are rejecting null hypothesis
-------------------------
Reviews:
-17.20042657460804
p-values:  0.0
 we are rejecting null hypothesis
-------------------------
Price:
1.4170805299182785
p-values:  0.0
 we are rejecting null hypothesis
-------------------------
```