Assignment - 1

Shantanu Tyagi (201801015)*
Dhirubhai Ambani Institute of Information & Communication Technology,
Gandhinagar, Gujarat 382007, India
CS306 - Data Analysis and Visualization

Question 1

outcomes (e)	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
Relative Probability Pr(e)	0.13	0.10	0.10	0.17	0.17	0.10	0.10	0.13

- E : More than one heads
- F : All coins are the same
- G : More than and equal to 2 tails
- H : Some coins different

(a) $P(E) = P(HHH, HHT, HTH, THH)$, Since these are mutually exclusive
$P(E) = P(HHH) + P(HHT) + P(HTH) + P(THH)$
$P(E) = 0.13 + 0.10 + 0.10 + 0.17 = 0.5$

$P(F) = P(HHH, TTT)$, Since these are mutually exclusive
$P(F) = P(HHH) + P(TTT)$
$P(F) = 0.13 + 0.13 = 0.26$

$P(E \cap F) = P(HHH)$
$P(E \cap F) = 0.13$

$P(E \cup F) = P(HHH, HHT, HTH, THH, TTT)$, Since these are mutually exclusive
$P(E \cup F) = P(HHH) + P(HHT) + P(HTH) + P(THH) + P(TTT)$
$P(E \cup F) = 0.13 + 0.10 + 0.10 + 0.17 + 0.13 = 0.63$

(b) We can directly use the addition rule, $P(E \cup F) = P(E) + P(F) - P(E \cap F)$. Putting the values calculated above we get, $P(E \cup F) = 0.5 + 0.26 - 0.13 = 0.63$, which gives us the same result.

(c) $P(G) = P(HTTTHTTTHTTT)$, Since these are mutually exclusive
$P(G) = P(HTT) + P(THT) + P(TTH) + P(TTT)$
$P(G) = 0.17 + 0.10 + 0.10 + 0.13 = 0.5$

$P(H) = 1 - \overline{P(H)}$
$P(H) = 1 - P(All\ coins\ same) = 1 - P(HHH, TTT)$, Since these are mutually exclusive
$P(H) = 1 - P(HHH) - P(TTT) = 1 - 0.13 - 0.13 = 0.74$

$P(G \cap H) = P(HTT, THT, TTH)$, Since these are mutually exclusive
$P(G \cap H) = P(HTT) + P(THT) + P(TTH)$
$P(G) = 0.17 + 0.10 + 0.10 = 0.37$

$P(G \cup H) = 1 - \overline{(G \cup H)}$
$P(G \cup H) = 1 - P(HHH)$
$P(G \cup H) = 0.13$

(d) We can check of G and H are independent if, $P(G \cap H) = P(G) \cdot P(H)$. Putting the values, we get $0.37 = 0.5 \cdot 0.74 = 0.37$. Hence G and H are independent events.

*Electronic address: 201801015@daiict.ac.in

Question 2 Let, $P(F)$ = Probability of faulty used car
$P(G)$ = Probability of good used car
$P(MF)$ = Probability that mechanic finds a car faulty
$P(MG)$ = Probability that mechanic finds a car good

We have,
$P(F) = 0.35$
$P(G) = 0.65$
$P(MF/F) = 0.94$
$P(MG/F) = 0.06$
$P(MG/G) = 0.88$
$P(MF/G) = 0.12$

$P(MF) = P(MF/F) \cdot P(F) + P(MF/G) \cdot P(G) = 0.407$
$P(MG) = P(MG/F) \cdot P(F) + P(MG/G) \cdot P(G) = 0.5932$

(a) Before getting advise, chances of getting a faulty car, i.e. P(F) = 0.35

(b) When the mechanic has certified the car as faulty, the chances of car actually being faulty, i.e. P(F/MF), can be found out using,

$P(F/MF) = \frac{P(F \cap MF)}{P(MF)}$

$P(F/MF) = \frac{P(MF/F) \cdot P(F)}{P(MF)}$

$P(F/MF) = \frac{0.94 \cdot 0.35}{0.407}$

$P(F/MF) = 0.8083$

(c) When the mechanic has certified the car as good, the chances of car actually being faulty, i.e. P(F/MG), can be found out using,

$P(F/MG) = \frac{P(F \cap MG)}{P(MG)}$

$P(F/MG) = \frac{P(MG/F) \cdot P(F)}{P(MG)}$

$P(F/MG) = \frac{0.06 \cdot 0.35}{0.5932}$

$P(F/MG) = 0.0354$

Question 3 $P(A)$ = Probability of disease A = $\frac{1}{20}$
$P(B)$ = Probability of disease B = $\frac{1}{10}$
$P(C)$ = Probability of no disease = 0
$P(H)$ = Probability of Headache

We have,
$P(H/A) = \frac{19}{20}$
$P(H/B) = \frac{1}{4}$
$P(H/C) = \frac{1}{10}$

(a) Given patient has complained of having a headache. In this case, probability of having disease A is given by $P(A/H)$
$P(A/H) = \frac{P(H/A) \cdot P(A)}{P(H/A) \cdot P(A) + P(H/B) \cdot P(B) + P(H/C) \cdot P(C)}$

Putting the values, we get,
$P(A/H) = 0.3015$

(b) Given patient has complained of having a headache. In this case, probability of having disease B is given by $P(B/H)$
$P(B/H) = \frac{P(H/B) \cdot P(A)}{P(H/A) \cdot P(A) + P(H/B) \cdot P(B) + P(H/C) \cdot P(C)}$

Putting the values, we get,
$P(B/H) = 0.1587$

(c) Given patient has complained of having a headache. In this case, probability of not having any disease is given by $P(C/H)$
$P(C/H) = \frac{P(H/C)\cdot P(C)}{P(H/A)\cdot P(A)+P(H/B)\cdot P(B)+P(H/C)\cdot P(C)}$

Putting the values, we get,
$P(B/H) = 0.5396$

(d) Since no patient has both the diseases simultaneously, we have the probability of having both the diseases given the patient complained of headache as 0.

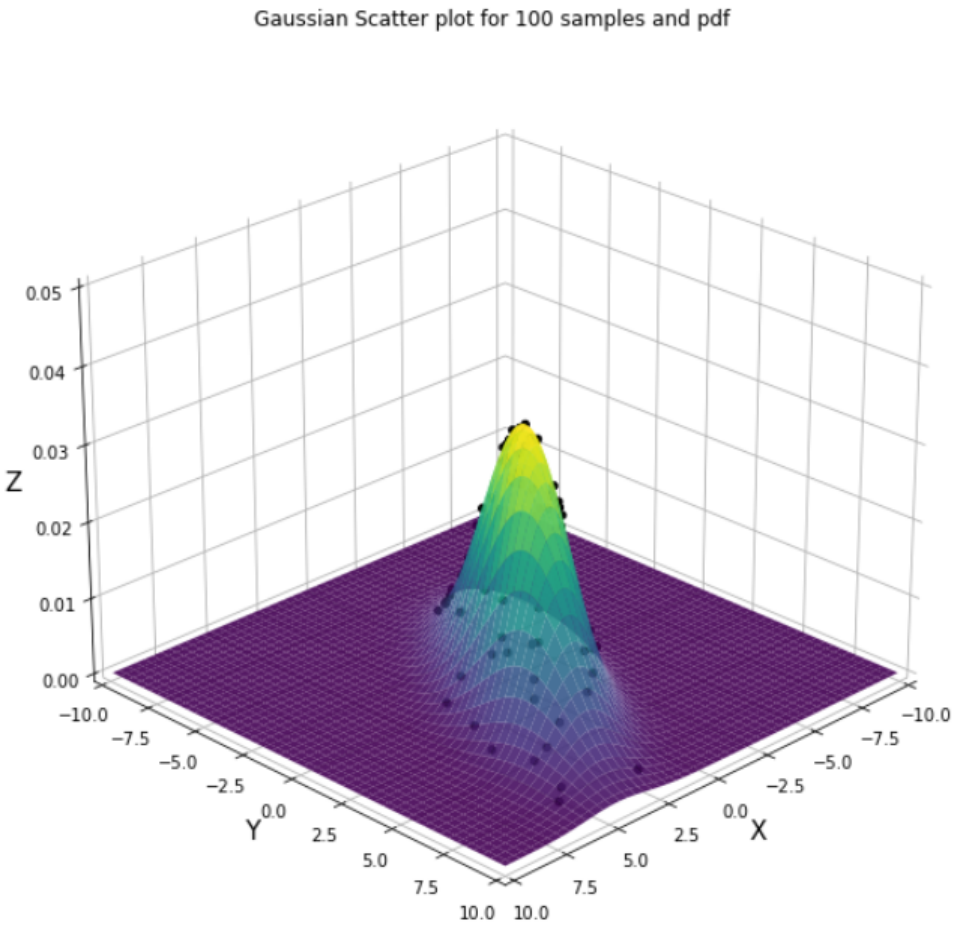
Question 4

(a) Please refer to the code since all the calculations have been done in the code. Mean men: 22.0
Mean women: 21.5
Standard deviation men: 11.627553482998906
Standard deviation women: 10.984079387914129
Ratio: 1.0585824330250924
95% interval: [-10.126793123757704, 11.126793123757704]
99% interval: [-14.05961502539322, 15.05961502539322]
Difference in mean: 0.5
Standard Error: 5.058161721416191

(b) We see that the value 0 in included in the confidence interval. This is the check for null hypothesis which in our case is true meaning that both men and women have similar income.

Question 5 Please refer to the code file since all the calculations have been done there.

(a) The 3D scatter plot of the data is given below.



(b) Sample Mean: [0.6550697038546436, 1.278693377837614]
Sample covariance: [[3.98993521 3.93271716] [3.93271716 8.8515139]]

(c) After doing a few Monte Carlo simulations and taking the average, we get the following results, Sample Mean: [1.00050645 2.02519283]
Sample covariance: [[3.90793726 3.79431866] [3.79431866 8.86311036]]
These value are more accurate and we can get better approximation if we increase N.

(d) On varying the trials(N), we get the following results,

	Mean	Co-variance	Mean RMSE	Cov RMSE
20	[0.994, 2.03]	[[3.956, 4.028], [4.028, 9.193]]	0.0216333	0.100937
40	[0.972, 1.986]	[[3.893, 3.825], [3.825, 8.787]]	0.0221359	0.171805
60	[1.03, 2.039]	[[4.033, 4.076], [4.076, 9.029]]	0.0347922	0.058056
80	[1.003, 2.013]	[[3.992, 3.986], [3.986, 8.991]]	0.00943398	0.0115866
100	[0.992, 1.989]	[[3.923, 3.912], [3.912, 8.986]]	0.00961769	0.0735068
200	[0.994, 1.979]	[[3.981, 3.956], [3.956, 8.849]]	0.0154434	0.0822101
300	[0.99, 1.979]	[[4.003, 3.997], [3.997, 8.98]]	0.0164469	0.010332
400	[1.0, 1.979]	[[4.002, 4.006], [4.006, 9.005]]	0.0148492	0.00502494
500	[0.988, 1.984]	[[3.994, 3.978], [3.978, 8.978]]	0.0141421	0.0192873

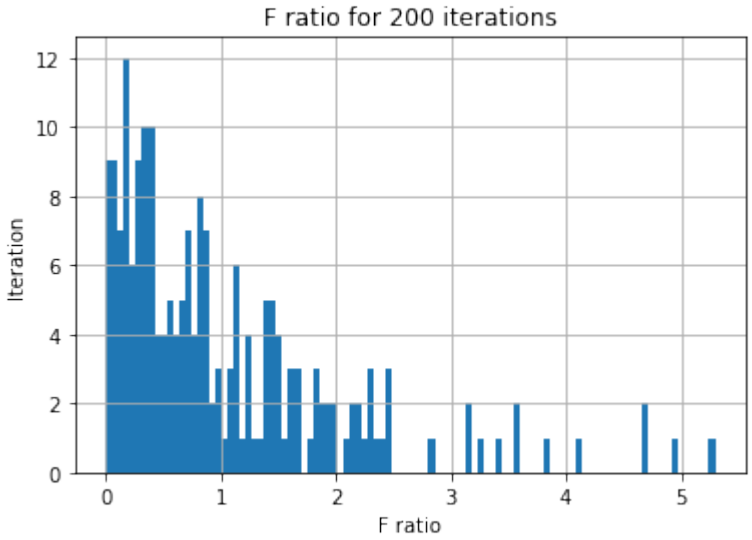
Question 6

(a) The ANOVA table calculated is shown below,

Source	Sum of Squares	Degress of Freedom	Mean Square	F-value
Between samples	511.044	2	255.522	2.0719
Within samples	3329.827	27	123.326	
Total	3,840.871	29		

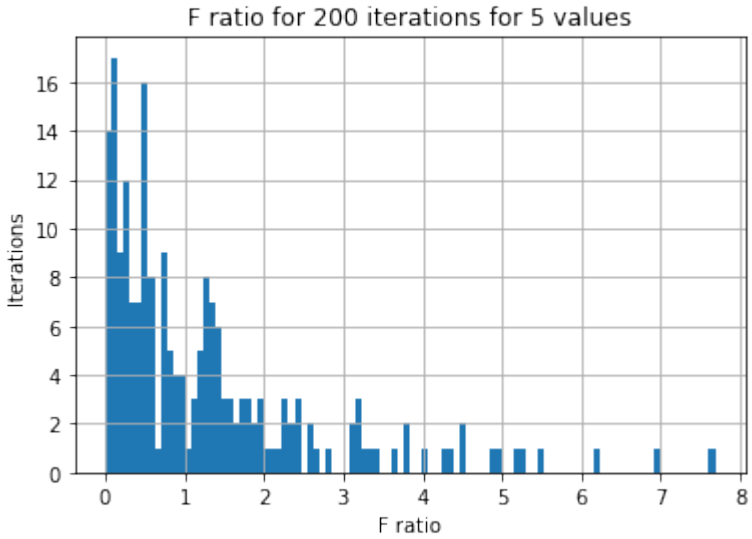
(b) After running the code, we get the following values, meanA: 60.71769969768949
meanB: 62.13156779622678
meanC: 64.27842260963371
overalMean: 62.37589670118333
SSB: 64.28918748518151
SSE: 4580.0125084983265
MSB: 32.14459374259076
MSE: 169.63009290734541
F: 0.18949817919482384

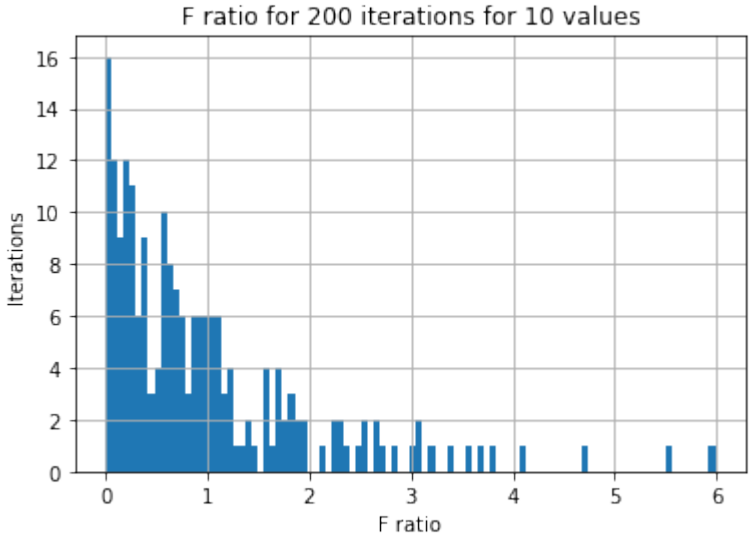
(c) Histogram visualisation for the resulting distribution of F ratio values taken over 200 runs is shown below.

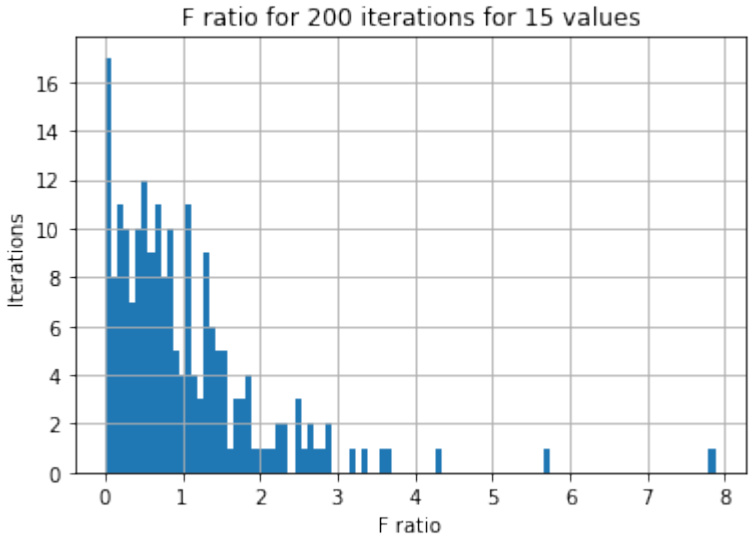


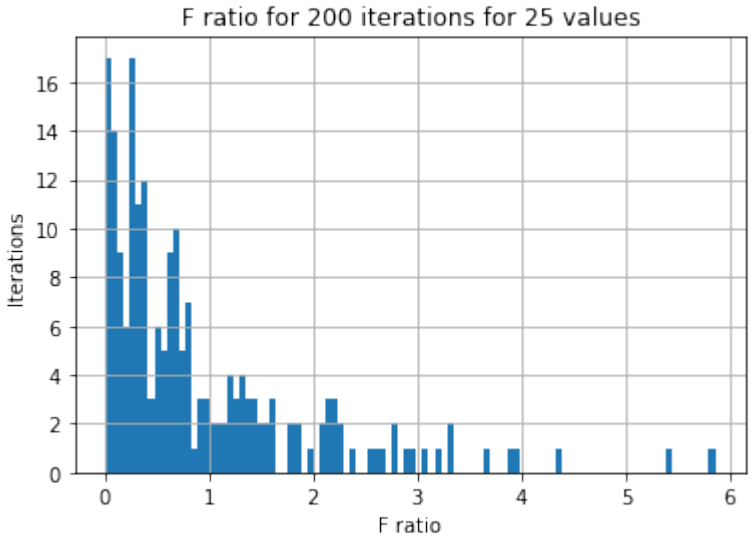
The proportion of F values that exceed $F_critical = 3.35$ when $\alpha = 0.05$ is 0.045. This value is very less and thus are from the same distribution since the hypothesis is accepted.

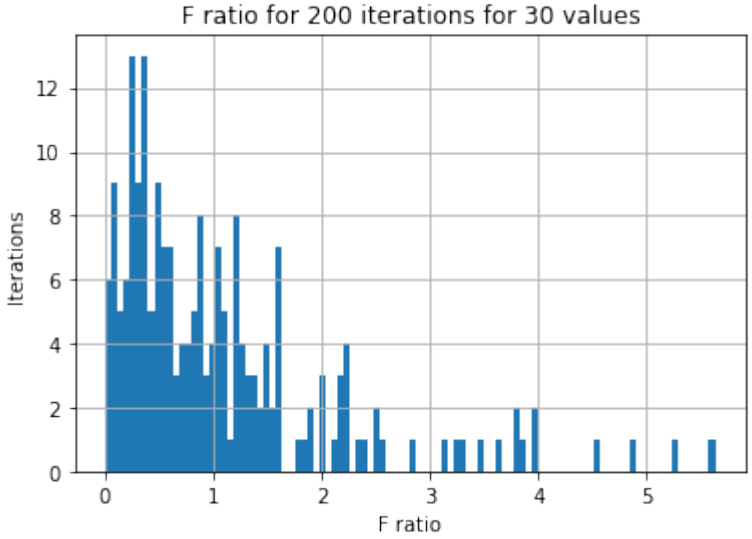
(d) By changing the number of observations, we get the following plots for the F statistic.







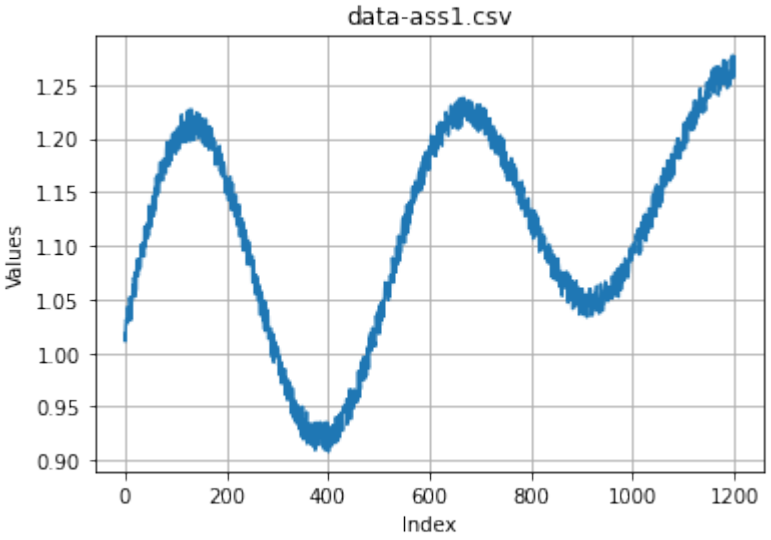




(e) In ANOVA test, we have as the significance level which is the probability of rejecting the null hypothesis when the result is true. It depicts the risk of assuming difference in the data even if no difference actually exists .This value if used to determine the threshhold value for F. The null hypothesis is accepted if the F values lie below this threshhold.

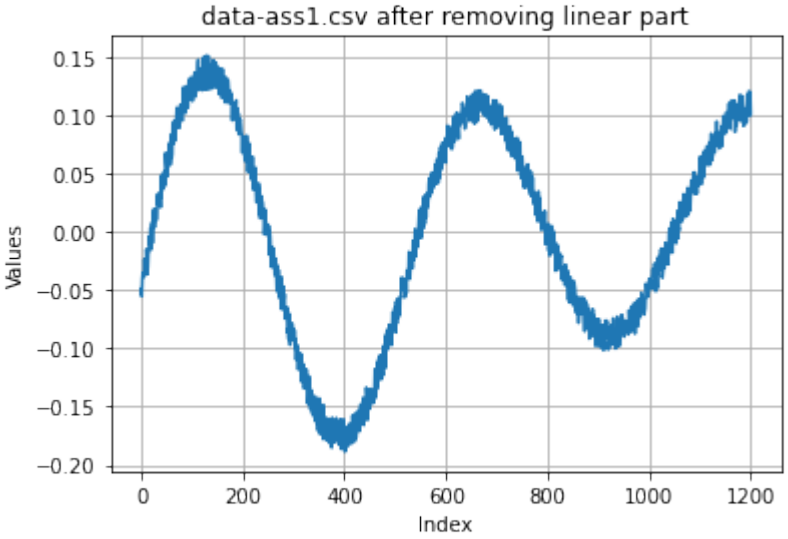
Question 7

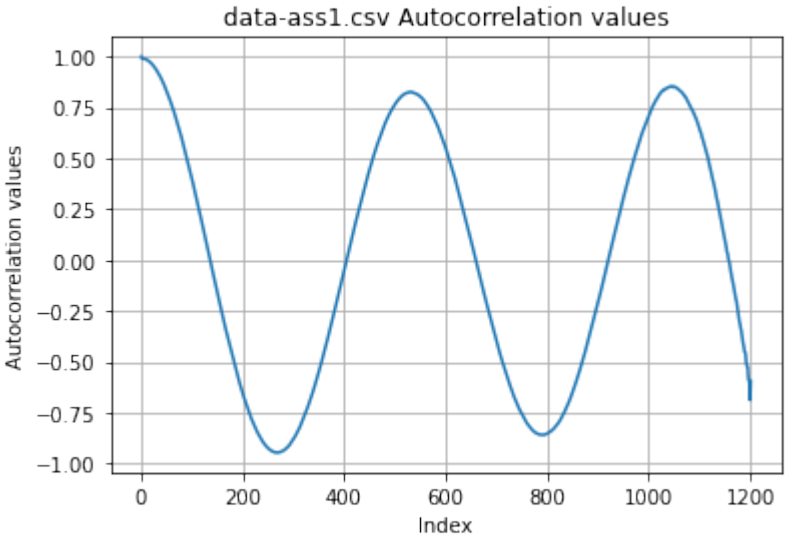
(a) The plot for the first dataset is given below.



The figure seems to be periodic globally but with local linear trends.

(b) We use linear regression to remove the linear trend in the data and check for periodicity. This is done by subtracting the values estimated by linear regression parameters. Next, we find the autocorrelation of the data. If this value if large it means that there is some periodicity in the data depending on the shift done during calculating autocorrelation, we can see which shift results in higher values. The difference in the peak values can help us find the period in the data.



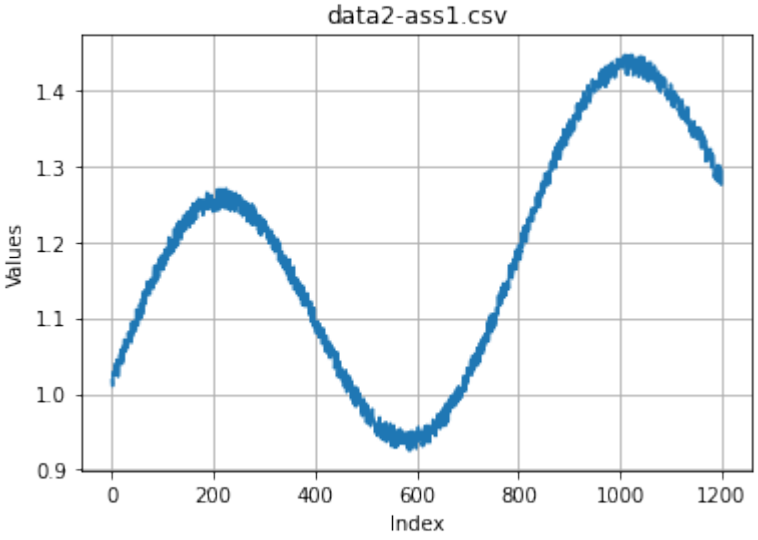


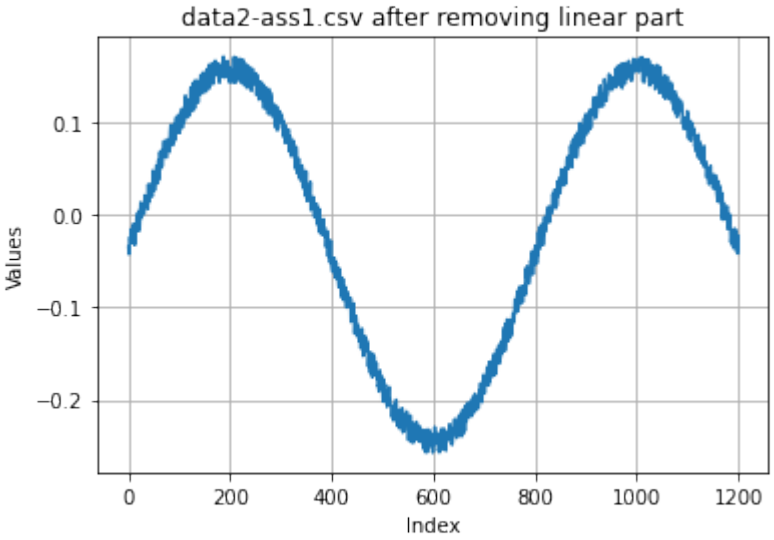
Peaks: [0, 3, 265, 269, 521, 528, 531, 533, 784, 786, 791, 1029, 1036, 1042, 1045, 1049] We can see two peaks, i.e. around 528(average) and around 1041(approx). Taking the difference, we get the period as 513(approx).

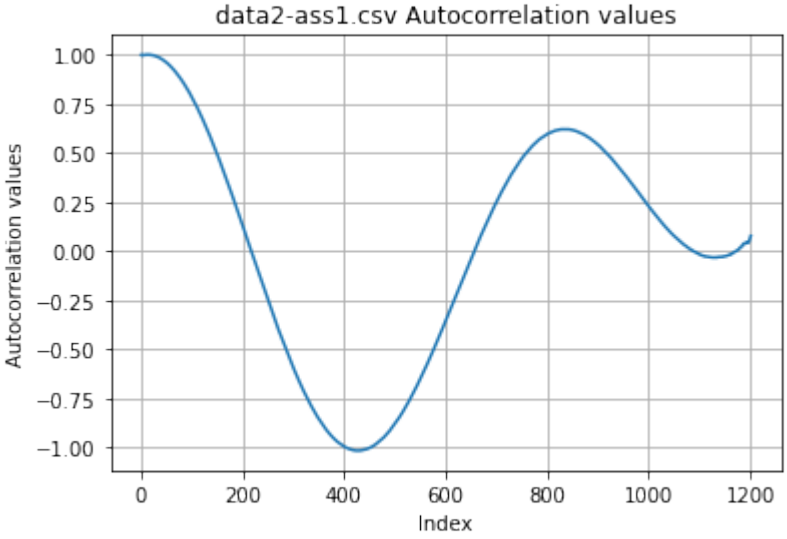
(c) The pseudo code for the same is,

```
df = read(data.csv)
y = df.values
plot(y)
linear_regress = LinearRegression().fit(index,y)
intercept = linear_regress.intercept
slope = linear_regress.slope
predicted = linear_regress.predict(df.index)
y = y - predicted
plt.plot(y)
y = flatten(y)
auto = autocorrelation(y)
plot(auto)
peak = peaks(auto)
period = period(peak)
```

(d) On repeating the same procedure on the second dataset, we get the following plots:







Peaks: [0, 11, 14, 428, 825, 829, 832, 835, 846, 1105, 1109, 1111, 1119, 1123, 1126, 1129, 1131, 1134, 1138, 1140, 1142, 1147, 1149, 1160, 1162, 1169, 1187, 1191, 1195] We can see two peaks, i.e. around 832(average) and around 1191(approx). Taking the difference, we get the period as 359(approx).

---

[1] https://www.analyticsvidhya.com/blog/2020/06/introduction-anova-statistics-data-science-covid-python/
[2] https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-233