ENGR_ALDA_FALL2021

GITHUB REPOSITORY - https://github.ncsu.edu/sjanard/engr-ALDA-fall2021-H12

Homework Team 12 - HW12

Sudharsan Janardhanan - sjanard
Sriram Sudharsan - ssudhar
Pradyumna Vemuri - pvemuri

---

1. (20 points) [Data Attributes] [Designed by Angela Zhang, Graded by Chengyuan Liu] Classify the following attributes as:

1) nominal, ordinal, interval, or ratio; and 2) as binary, discrete, or continuous. Some cases may have more than one interpretation, so briefly justify your answer if you think there may be some ambiguity.

(a) (1 point) The pH of water that has been measured using the pH scale. - **Interval, Continuous.**

In the pH scale, 3 means acidic and 6 would mean slightly acidic. Since pH is a Interval value, ratios are not meaningful. But a 3 is not half the value of 6 in terms of acidity.

(b) (1 point) Population counts in counties of North Carolina. - **Ratio, Discrete**

(c) (1 point) Active Duty Enlisted Basic Military pay grade scale e.g. E-1, E-2, ..., E-9(for simplicity of the exercise, do not factor in years of service in your answer). - **Ordinal, Discrete**

(d) (1 point) The amount of Calories found on a nutrition label. - **Ratio, Continuous**

(e) (1 point) The distance traveled on a road trip measured in miles. - **Ratio, Continuous**

(f) (1 point) How much time until midnight? **- Ratio, Continuous.**

Any time based attribute is generally perceived as an Interval type, but at midnight the time resets. Since there is a proper zero point in the scale, here it can be a Ratio.

(g) (1 point) Percentage of daily vitamin C intake from a glass of orange juice. - **Ratio, Continuous**

(h) (1 point) Existence and non-existence of cancerous tumors. - **Nominal, Binary**

(i) (1 point) Annual income (US currency). - **Ratio, Discrete**

(j) (1 point) An angle measured in radians between 0 and 2. - **Ratio, Continuous**

Since angles can have a zero value and ratios make sense, so the attribute is a ratio.

(k) (1 point) Number of fishes in an aquarium. - **Ratio, Discrete**

(l) (1 point) Required or Not Required. - **Nominal, Binary**

(m) (1 point) Level of Pain during Injury diagnosis (0: absent, 1: mild, 2: moderate, 3: severe, 4: incapacitating). - **Ordinal, Discrete**

(n) (1 point) Circumference of a pizza. - **Ratio, Continuous**

(o) (1 point) Product Satisfactory Rating (5 - Excellent, 4 - Good, 3 - Fair, 2 - Poor, 1- Horrible). - **Ordinal, Discrete**

(p) (1 point) Temperature in degrees Fahrenheit. - **Interval, Continuous**

(q) (1 point) A person's weight measured in pounds. - **Ratio, Continuous**

(r) (1 point) Correct or Incorrect. - **Nominal, Binary**

(s) (1 point) Number of pieces remaining to complete a jigsaw puzzle. - **Ratio, Discrete**

(t) (1 point) Day of the month. - **Interval, Discrete.**

This depends on the way the ‘day of the month’ is perceived. If we mean Monday, Tuesday, etc, then the attribute would be nominal. But in general if the date is considered, it will be an Interval attribute.

Source code for Question 2 can be found in the python file 2.py in the following github repository under H1 directory.

https://github.ncsu.edu/sjanard/engr-ALDA-fall2021-H12

2. (15 points) [Matrix Operations] [Ge Gao] Write code in Python to perform each of the following tasks, please report your output and relevant code in the document file, and also include your code file (ends with extension .py) in the .zip file.

(a) (1 point) Generate a 5*5 identify matrix A.

```
[[1, 0, 0, 0, 0], [0, 1, 0, 0, 0], [0, 0, 1, 0, 0], [0, 0, 0, 1, 0], [0, 0, 0, 0, 1]]  is the identity matrix
```

(b) (1 point) Change all elements in the 3rd column of A to 2.

```
[[1, 0, 2, 0, 0], [0, 1, 2, 0, 0], [0, 0, 2, 0, 0], [0, 0, 2, 1, 0], [0, 0, 2, 0, 1]]  after changing third column to 2
```

(c) (1 point) Sum of all elements in the matrix (use ONE "for/while loop"). (d) (1 point) Transpose the matrix A (A=A^T).

```
sum of all the elements in the matrix  14
```

(d) (1 point) Transpose the matrix A (A=AT).

```
[[1, 0, 0, 0, 0], [0, 1, 0, 0, 0], [2, 2, 2, 2, 2], [0, 0, 0, 1, 0], [0, 0, 0, 0, 1]]  is the transposed matrix
```

(e) (2 points) Calculate the sum of the 3rd row, the sum of the diagonal and the sum of the 2nd column in matrix A, respectively (your answer should be three numbers).

```
10
6
3
```

(f) (1 point) Generate a 5*5 matrix B following standard normal distribution.

```
[[ 0.49671415 -0.1382643   0.64768854  1.52302986 -0.23415337]
 [-0.23413696  1.57921282  0.76743473 -0.46947439  0.54256004]
 [-0.46341769 -0.46572975  0.24196227 -1.91328024 -1.72491783]
 [-0.56228753 -1.01283112  0.31424733 -0.90802408 -1.4123037 ]
 [ 1.46564877 -0.2257763   0.0675282  -1.42474819 -0.54438272]]
```

(g) (2 points) From A and B, using matrix operations to get a new 2*5 matrix C such that, the first row of C is equal to the 1st row of B minus the 2nd row of A, the second row of C is equal to the sum of the 4th row of A and the 5th row of B.

```
[[ 0.49671415 -1.1382643   0.64768854  1.52302986 -0.23415337]
 [ 1.46564877 -0.2257763   0.0675282  -0.42474819 -0.54438272]]  is the resultant matrix
```

(h) (2 points) From C, using ONE matrix operation to get a new matrix D such that, the first column of D is equal to the first column of C times 2, the second column of D is equal to the second column of C times 3 and so on.

```
[array([0.99342831, 2.93129754]), array([-3.4147929, -0.6773289]), array([2.59075415, 0.27011282]), array([ 7.61514928, -2.12374093]), ar
ray([-1.40492025, -3.26629635])]
```

(i) (2 points) X = [1, 3, 5, 7]^T, Y = [4, 3, 2, 1]^T, Z = [2, 4, 6, 8]^T. Compute the covariance matrix of X, Y, and Z. Then compute the Pearson correlation coefficients between X and Y.

```
[[ 6.66666667 -3.33333333  6.66666667]
 [-3.33333333  1.66666667 -3.33333333]
 [ 6.66666667 -3.33333333  6.66666667]]
array([[ 1., -1.],
       [-1.,  1.]])
```

(j) (2 points) Verify the equation: $\overline{x^2} = (\bar{x}^2 + (x))$ using x = [20, 1, 3, 5, 7, 9, 14]^T when (python library *math* is allowed):

i. $\sigma(x)$ is the population standard deviation. Show your work.

ii. $\sigma(x)$ is the sample standard deviation. Show your work.

In the equation:

LHS of the equation = mean of square of x

RHS of the equation = sum of square of mean of x and variance of x

```
LHS Mean is :  108.71428571428571
RHS Mean for Population Standard Deviation 108.71428571428572
RHS Mean for Sampling Standard Deviation 114.99319727891157
```

Hence, we observe that the equation mentioned above holds true for the population standard deviation.

Source code for Question 3 can be found in the python file 3.py in the following Github repository.

https://github.ncsu.edu/sjanard/engr-ALDA-fall2021-H12

3. **(24 points) [Data Visualization] [Designed by John Wesley Hostetter, Graded by Chengyuan Liu] In this question, please summarize and explore data in the provided file "data\wine.csv". This data is the "Wine Data Set" and is a famous database donated by Stefan Aeberhard to the UCI Machine Learning Repository.**

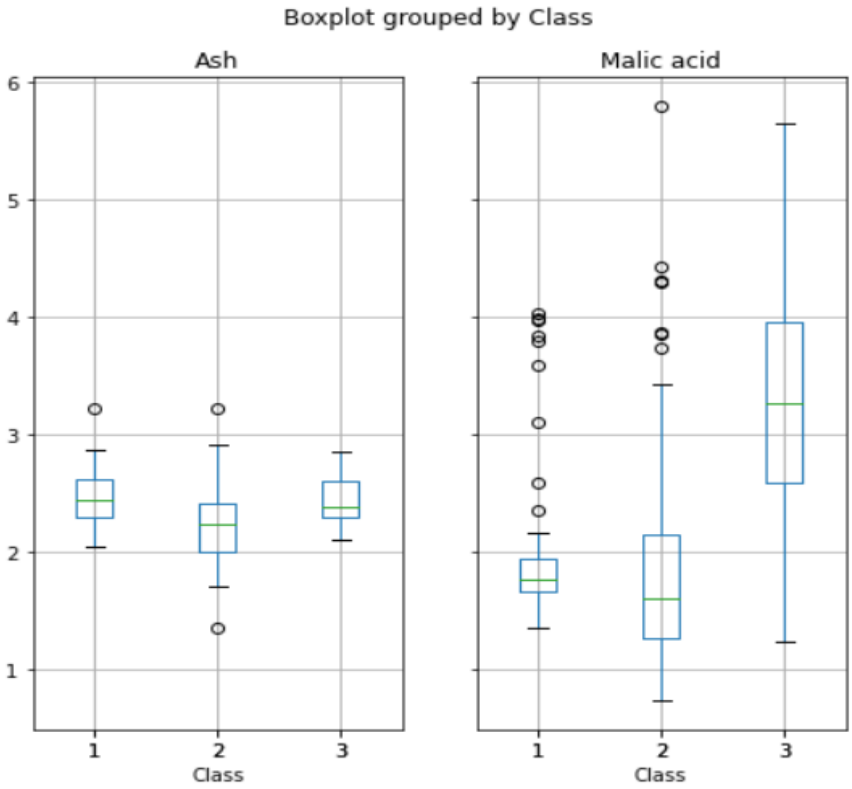
   Write code in Python to perform the following tasks. Please *report your output and relevant code* in the document file, and also include your code file (ends with extension .py) in the .zip file.

   (a) (4 points) Compute the mean, median, standard deviation, range, 25th percentiles, 50th percentiles, 75th percentiles for the following attributes: *Alcohol*, *Malic acid*, *Ash*, *Alcalinity Of Ash*.

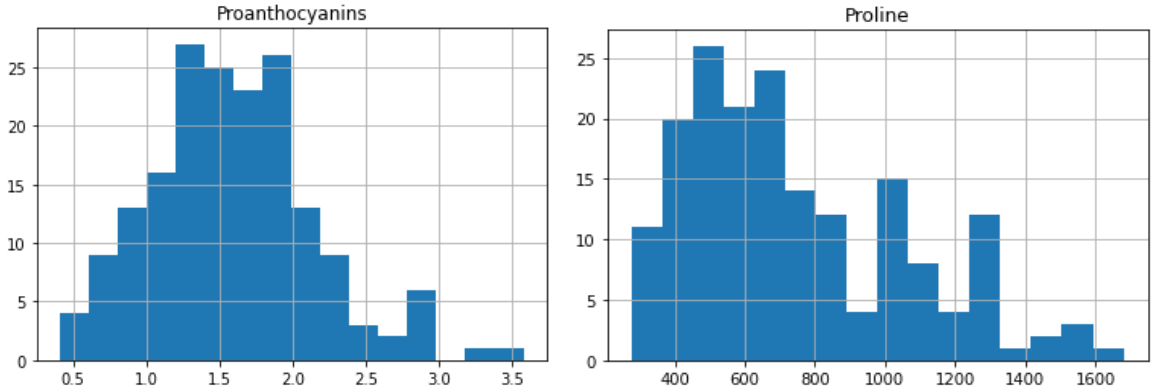
   Following are the values for Alcohol, Malic acid, Ash & Alcalinity of Ash respectively.

```
Mean =  13.000617977528083 Median =  13.05 Standard Deviation = 0.8118265380058577 Range =  3.8000000000000007
25th percentile = 12.362499999999999 50th percentile = 13.05 75th percentile 13.6775

Mean =  2.336348314606741 Median =  1.8650000000000002 Standard Deviation = 1.1171460976144627 Range =  5.06
25th percentile = 1.6025000000000003 50th percentile = 1.8650000000000002 75th percentile 3.0825

Mean =  2.3665168539325854 Median =  2.36 Standard Deviation = 0.2743440090608148 Range =  1.8699999999999999
25th percentile = 2.21 50th percentile = 2.36 75th percentile 2.5575

Mean =  19.49494382022472 Median =  19.5 Standard Deviation = 3.3395637671735052 Range =  19.4
25th percentile = 17.2 50th percentile = 19.5 75th percentile 21.5
```

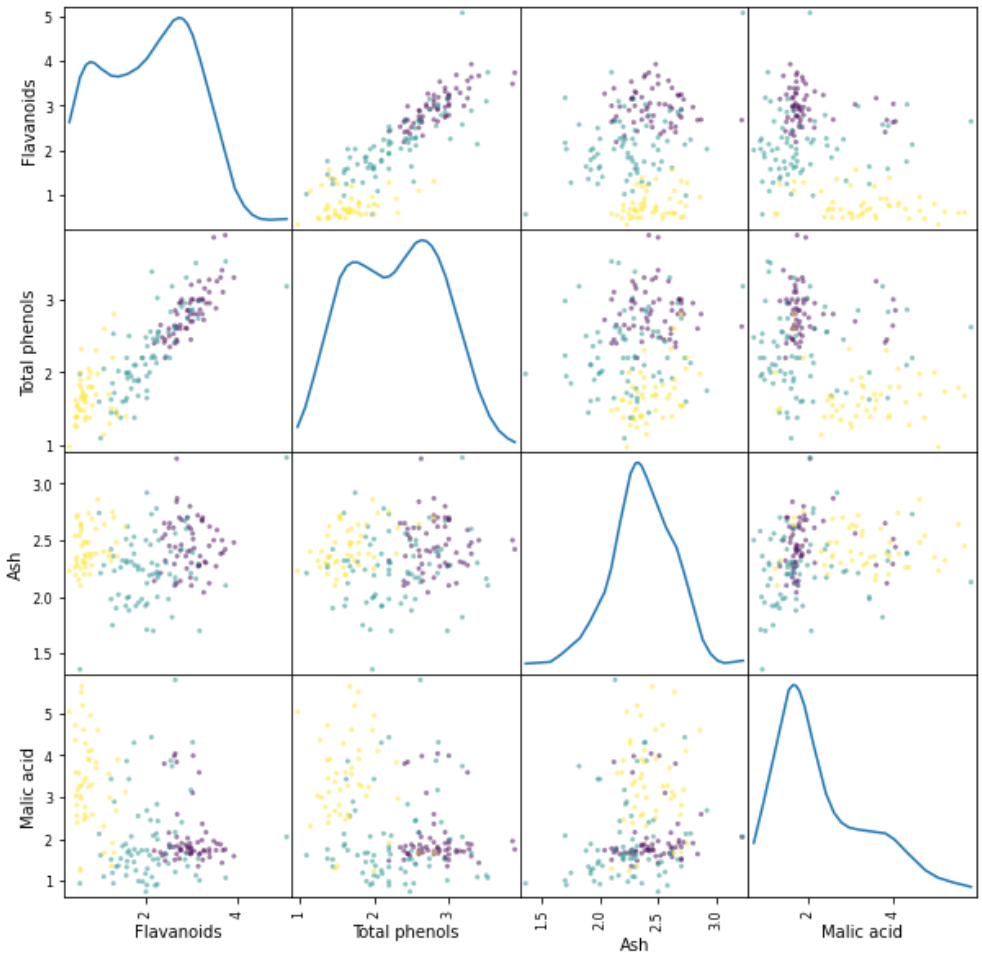
(b) (3 points) Make a box-and-whisker plot for the attributes *Ash* and *Malic acid* where they are grouped by the *class* label. Be sure to include a title for each plot of what feature is being described.



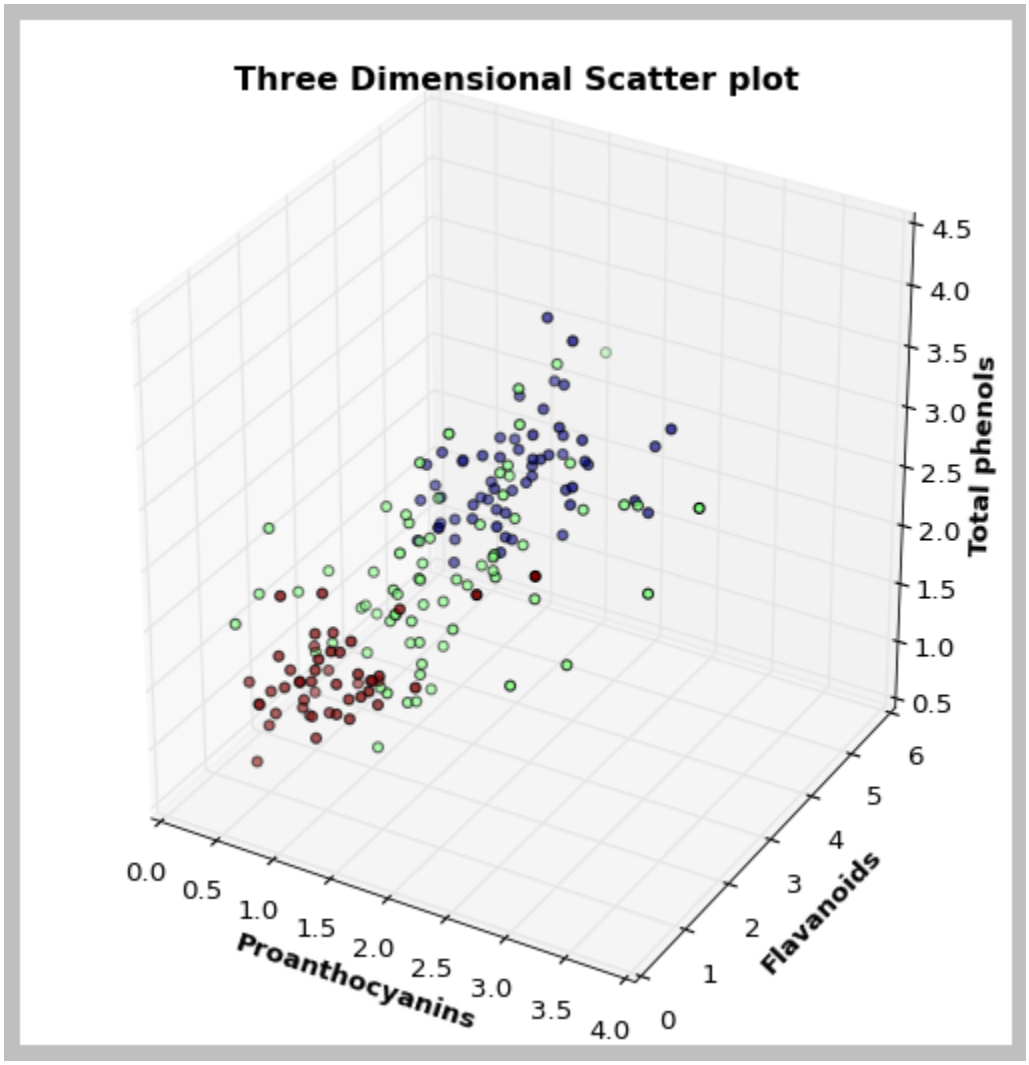
(c) (4 points) Create histogram plot using 16 bins for the two features *Proanthocyanins* and *Proline*, respectively.



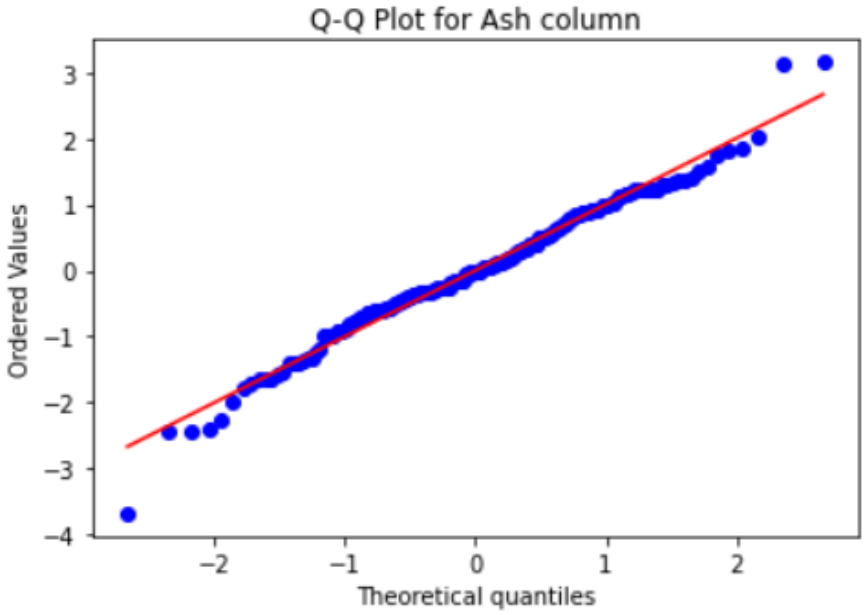
(d) (4 points) Create a scatter matrix of the data. Include only the following features: *Flavanoids*, *Total phenols*, *Ash*, *Malic acid*, but use the *class* attribute to change the color of the data points (for convenience, you may use a library for this). For the diagonal of the scatter matrix, plot the kernel density estimation (KDE).

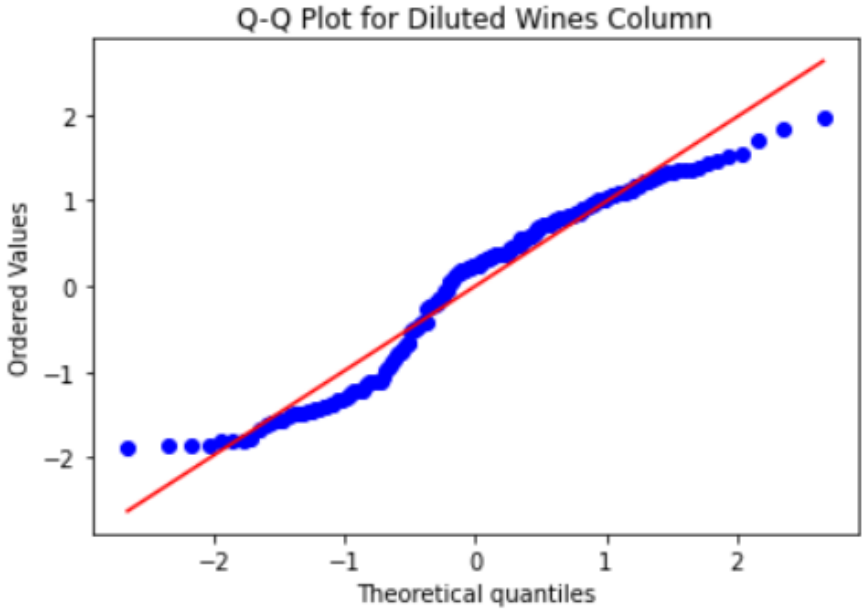


(e) (5 points) Now, write code to produce a three-dimensional scatter plot using the *Proanthocyanins*, *Flavanoids* and *Total phenols* as dimensions, and color the data points according to the *class* attribute.



(f) (4 points) The quantile-quantile plot can be used for comparing the distribution of data against the normal distribution. Create a quantile-quantile plot for the two features *Ash* and *OD280/OD315 of diluted wines*, respectively. Give a brief analysis for the two plots.





The Q-Q plot while plotted against the same attribute provides an insight into the distribution of values in that attribute. The reference line indicates perfect normal distribution and any variation from it would mean that the data is skewed. The Q-Q plot for Ash has values very close to the line with an exception of a few outliers as compared to the Diluted Wines attribute where the data is a bit more varied. This would suggest that the values in the Ash attribute are more normally distributed than that of Diluted Wines which seems to be more skewed.

4. (16 points) [Short Answer Questions] [Angela Zhang] Please read Chapters 2 in Tan et al., textbook and review lecture notes to answer the following questions:

(a) (10 points) General Short Answer

i. (3 points) Which distance metric would best describe this: How far away is the agent from their goal in a GridWorld environment? Justify your answer.

In the grid world environment, an agent is allowed only four directions of movement (North, South, East and West). The agent is expected to get from one point to another moving in these directions, one grid square at a time. To calculate how far an agent is from their goal, the **Manhattan Distance (Minkowski Distance, r=1)** would be the best option. This calculates the distance between two real valued vectors. Manhattan Distance can be visualized in a coordinate system as a line that traverses along the x and y axes of the grid.

ii. (3 points) What is the definition of covariance? If variables A and B have a covariance of -345 while variables B and C have a covariance of 20. What claims Can you draw? Justify your answer.

In statistics, covariance is used to determine the relationship between two random variables. It quantifies how each of the dimensions varies from their respective mean(s). However, It should be noted that as it doesn't employ a standardized unit of measurement. It may not be possible to nail down the degree of change but however, it is possible to make conclusions based on the sign of covariance alone and not its absolute value.

- Positive correlation implies that the relationship between the two variables moves in the same direction (in tandem). When one variable increases, the other also increases. Similarly, when one variable decreases, the other also decreases.

- A negative correlation implies that an increase/decrease in the first variable causes a decrease/increase (opposite effect) in the second variable.

- In the first case, where A and B have a covariance of -345, it can be inferred that for an increase in the value of A, there is a decrease in the value of B (or vice versa).

- However, In the second case, the positive covariance (+20) of B and C implies that a change in B produces a proportional change in C ie. an increase in B, causes an increase in C (and vice versa).

iii. (4 points) Provide a scenario in which you might encounter duplicate data. What could have caused the data to be duplicated? How would it be detected? Provide a solution to resolve the duplication, and state the pros/cons.

- Duplicate data is commonly encountered while posting advertising mails. For example, there may be multiple advertising mails delivered from an organization to the same address for the same person. This may be due to multiple enrollment by the same person using different credentials. In particular, this type of duplication is more pronounced in countries without a defined address convention.
- Simple querying on dataset can help identify and group duplicate data or Jaccard Similarity measure which is used generally for measuring similarity between two data objects can be scaled such that points in the dataset that are too similar can be identified as duplicates.
- The duplication can be resolved by comparing the duplicate addresses and matching them with a unique identifier. For example, in the case of a wireless network provider, all duplicate addresses can be mapped to a customerID (unique identifier). Once the unique identifier is established and addresses are mapped, the duplicate entries can be removed. However, this solution fails to consider a situation where one individual pays for two subscriptions at two locations.

(b) (6 points) Noise and Outliers

i. (2 points) In your own words, explain what noise is. Can noise ever be desirable? If so, give an example when it is desirable. If not, provide an explanation why not.

- Although algorithms are robust enough to handle erroneous data, most of the undesirable and useless data called **noise** needs to be removed or processed from the dataset.

- Noise occurs as a consequence of systematic errors or random measurement errors. It can either be a distortion of the attribute values of the existing objects or the addition of data objects that do not represent the rest of the data. Noise makes it hard for algorithms to detect patterns in data.

- Even though noise is often conceived as meaningless data, sometimes noise can be desirable.

  In Reinforced Learning, an agent in an environment performs various actions. Based on the action performed the agent receives feedback, from which it learns. The learning process depends on the agent's exploratory behavior, and introducing noise into the action space or the agent's parameters can result in a much more consistent exploration.

Another situation where noise may be desirable is in Image Processing. While training neural networks, noise can be added to our image for data augmentation purposes. This enables the neural network to train on more data. This approach is commonly used when there is insufficient amount of data. The types of noise added to the images are:

- Speckle Noise
- Gaussian Noise
- Salt and Pepper Noise

ii. (2 points) In your own words, explain what is an outlier? How could outliers be detected? How do outliers differentiate from noise?

- Outliers are data objects that do not fit in or appear analogous to the rest of the objects in the dataset. When the data points are represented in a multidimensional space, these outliers lie far away from the rest of the data.
- Outliers might have attribute values that are unusual compared to the rest of the data but are still valid and useful data.
- Outliers have a huge impact on popularly used statistical tools like mean, variance, and standard deviation. Therefore, a Z-score is used to quantitatively analyze attribute values to isolate outliers. On the other hand plots, histograms, and box plots can be used to identify outliers in the data.
- Outliers are still part of the dataset, they may or may not be removed depending on the algorithm being used to analyze the data. Outliers can be useful in applications like fraud detection, spam filters, medical analysis, etc. Noise on the other hand needs to be eliminated from the data during the pre-processing stage, as it may make it difficult for the algorithms to detect meaningful patterns in the data.

iii. (2 points) While conducting preliminary data exploration, you noticed that some patient's Electronic Healthcare Records (EHRs) had anomalous values that caught your attention. Specifically, you noticed that these patients' electrocardiogram results were significantly different from typical electrocardiogram behavior, but noticed that all these patients were in the most recent week of the collected data. You consult with your research partner, the oncologist, about your findings. Upon inspecting the electrocardiogram (EKG) machine that was used to collect the data, the oncologist discovers a fault within the machine's operation that led to corrupt electrocardiogram results. Unfortunately, the results are not salvageable, as the corruption appears to occur randomly throughout the electrocardiogram readings. Were the corrupted electrocardiogram results an example of noisy data or were they outliers in the data? Briefly justify your answer.

Given that the machine was faulty, noise is the likelier result. Outliers are data points that vary slightly from the regular data but are expected. In case the data produced by the machine were just outliers, the right data should have been salvageable. Since it's given that the data was unsalvageable, the only cause for that could be noise as noise taints the data and makes it unusable. Therefore we can confirm that the corrupted data was the result of noise and not outliers.

---

5) (9 points) [Sampling] [John Wesley Hostetter] Answer the following questions:

(a) (3 points) Exploration of the environment is a key principle to learning optimal actions with online reinforcement learning (RL). Imagine the RL agent reaches a state for which it has no prior knowledge to assist it in its decision making on what action it should take. For simplicity, assume that the environment is \safe" to explore within e.g. no human lives will be harmed or lost if the agent were to err. Let A denote the finite set of available actions to the agent when it reaches this novel state. Which sampling method would be appropriate and why? Briefly justify your answer.

Random Sampling with replacement

In the above case, when the agent reaches a novel state, only a finite set of actions are available. In addition, no information about previous actions are given. We need to select a sample set that has an equal representation of all available actions. As it is mentioned in this case that there are no harmful implications when the agent makes an error (considered to be safe), there is no effective need for taking the safer approach of progressive sampling. It is these reasons that encourage us to employ random sampling. Doing so gives the agent the option to return to the same state and explore alternate actions. Another compelling reason to use random sampling in this case would be the reduction of time and space complexity as the actions are chosen at random and wouldn't be done in an interactive manner. This would render our method extremely efficient. Apart from this, the agent doesn't have to consider replacing the action back into the available actions as the objective would be to apply as many actions as possible and learn from it efficiently. Hence, random sampling without replacement would be the best option.

(b) (3 points) Assume you have a large collection of data (e.g. billions of logged human-computer interactions), and want to apply a machine learning algorithm to obtain a predictive model for some mundane task. Despite the size of the data, the data itself is not very complicated, and the domain it describes is rather trivial. In addition, the algorithm you would like to use has a time complexity of O(n3), and space complexity of O(n2), where n is the number of data observations. For the purpose of this exercise, assume your supervisor expects your algorithm to have at least 90% sensitivity on the validation data set. Which sampling method would be appropriate and why? Briefly justify your answer.

Progressive sampling

This approach begins with a small sample size, and then gradually increases until an adequate sample size is obtained. Despite the model's accuracy increasing as the size of the sample increases, at the level-off point the growth in accuracy stops. It is at

this levelling-off point that the sampling process ceases. In this case, the sample size can be increased until the 90% sensitivity threshold is reached.

(c) (3 points) An oncologist approaches your research laboratory to collaborate together on a cancer detection system. This system, when provided with a patient's electronic health record (EHR), would then output the predicted probability that the patient has cancer (irrespective of the type of cancer). Fortunately, this type of cancer is rare, among the 793,000 labeled EHRs, only 1,560 have cancer, where the rest do not. You are required to create three sets of balanced data from this collection. Which sampling method would be appropriate and why? Briefly justify your answer.

Stratified sampling

This type of sampling is used when there are different kinds of objects with a different number of objects, that is the data is imbalanced. This hinders the objects with a lesser frequency to be sufficiently represented. Accordingly, a fixed number of objects are extracted from each representative group. There are two possible variations to this approach:

- Despite the different sizes of each group, an equal number of objects are selected from every group.

- The total count of objects selected from each group is proportional to the size of that group.

Inorder to create the three sets of balanced data, equal number of labeled EHRs are extracted from both the positive and negative cases for every set. This ensures that the cancer detection system has an equal representation of cancer positive and cancer negative EHRs.

---

6) (16 points) Data Transformation. [Created by Ge Gao and Graded by Chengyuan Liu]

(a) Please identify the appropriate data transformation methods for the following situations. Give a brief description about your answers:

i. (4 points) To learn a model to predict whether a patient suffers from organ failure, we need to consider two state features: heart rates and body temperature. Heart rates are distributed between 35 and 200 bpm (mean = 88, standard deviation = 18), and body temperatures are ranged from 84 to 112 (mean =98.6, standard deviation = 0.95).

1) For each feature, apply normalization(transformed data has: x0 2 [0; 1]) and calculate the new mean and new standard deviation of the normalized feature. Compare their means and standard deviations.

Mean = μ
Standard Deviation = S
Min value = X_{min}
Max value = X_{max}

New Mean = μ_n
New Standard Deviation = S_n

$\mu = \Sigma X / n = (X1 + X2 + ... + Xn) / n$

$X1' = (X1 - X_{min}) / (X_{max} - X_{min})$ ---------- 0)

$\mu_n = (X1' + X2' + ... + Xn') / n = ((X1 + X2 + ...+Xn) - n*X_{min}) / n * (X_{max} - X_{min})$
$= (n*\mu - n*X_{min}) / n * (X_{max} - X_{min}) = (\mu - X_{min}) / (X_{max} - X_{min})$ ---------- 1)

$S^2 = \Sigma(X - \mu)^2 / N$

$S_n^2 = ((X1' - \mu_n)^2 + ...+ (Xn' - \mu_n)^2) / N$

Substitute values from 0) and 1)

$S_n^2 = ((X1 - \mu)^2 + ...+ (Xn - \mu)^2) / N * (X_{max} - X_{min})^2$

$S_n = (S) / (X_{max} - X_{min})$

So based on the above derivation, for Min-Max Normalization:

$\mu_n = (\mu - X_{min}) / (X_{max} - X_{min})$

$S_n = (S) / (X_{max} - X_{min})$

For features Heart Rates and Body Temperatures:

- <u>Heart Rates:</u>

  $\mu_n = (\mu - X_{min}) / (X_{max} - X_{min})$

  $\mu_n = (88 - 35) / (200 - 35) = 53 / 165 =$ **0.321** (Normalized Mean)
  Original Mean μ =88

  $S_n = (S) / (X_{max} - X_{min})$

  $S_n = (18) / (200 - 35) = 18 / 165 =$ **0.109** (Normalized Standard Deviation)
  Standard Deviation S = 18

- <u>Body Temperature:</u>

  $\mu_n = (\mu - X_{min}) / (X_{max} - X_{min})$

  $\mu_n = (98.6 - 84) / (112 - 84) = 14.6 / 28 =$ **0.521** (Normalized Mean)

  Original Mean μ : 98.6

  $S_n = (S) / (X_{max} - X_{min})$

  $S_n = (0.95) / (112 - 84) = 0.95 / 28 =$ **0.034** (Normalized Standard Deviation)

  Original Standard Deviation : 0.95

For features Heart Rates and Body Temperatures:

And 2) for each feature, apply standardization to it and show the range of transformed data and compare their ranges.

Standardization using formula: $X_N = (X - \mu) / S$

- Heart Rates: Heart rates are distributed between 35 and 200 bpm. Mean is 88 and Standard Deviation is 18.

  New Range is: (X_{new_max} - X_{new_min})

  X_{new_min} = (X - μ) / S = (35 - 88) / 18 = -53 / 18 = **-2.94**.

  X_{new_max} = (X - μ) / S = (200 - 88) / 18 = 112 / 18 = **6.22**.

  Old Range = 200 - 35 = **165**

  New Range = 6.22 + 2.94 = **9.16**

- Body Temperature: Body Temperatures are distributed between 84 and 112 bpm. Mean is 98.6 and Standard Deviation is 0.95.

  New Range is: (X_{new_max} - X_{new_min})

  X_{new_min} = (X - μ) / S = (84 - 98.6) / 0.95 = -14.6 / 0.95 = -**15.36**.

  X_{new_max} = (X - μ) / S = (112 - 98.6) / 0.95 = 13.4 / 0.95 = **14.10**.

  Old Range = 112 - 84 = **28**

  New Range = 14.10 + 15.36 = **29.46**

ii) (4 points) During the design of an artificial neural network, we sometimes need to transform a variable x that has a range of (-∞,∞) to an open set z ∈ (-1, 1). Note that z monotonically increases as x increases in this transformation. Please specify a proper function for such transformation.

"Activation functions" can be used to determine the output of an Artificial Neural Network and map the output to values within the range of 0 to 1, or between -1 to 1 depending on the activation function being used.

To transform a variable x input value to a range of (-1,1), we can use **Hyperbolic tangent or Tanh** as the activation function.

(b) In natural language processing (NLP), there are diverse ways to represent words such as one-hot encoding, bag of words, TF*IDF, and distributed word representations. In one hot encoding, a bit vector whose length is the size of the vocabulary of words is created, where only the associated word bit is on (i.e., 1) while all other bits are 0 (i.e., 0). Here is a toy example: suppose there is a 5-dimensional feature vector to represent a vocabulary of ve words: [king, queen, man, woman, power]. In this case, `king' is encoded into [1,0,0,0,0], `queen' is encoded into [0,1,0,0,0], etc. Due to the nature of this representation, the feature vector encodes the vocabulary of a sentence where all words are equally distant. On the other hand, in distributed word vectors, a real-valued vector whose length is defined by some common properties of words is created, then each word can be represented as a linear combination of the defined properties. Using the toy example above, given a 3-dimensional feature vector of [man, woman, power] as the common properties, then words such as `king', `queen', `man', and `woman' could be encoded into [0.98,0.1, 0.8], [0, 0.99, 0.85], [0.9, 0, 0.5], and [0, 0.97, 0.5], respectively. In this case, if you subtract a vector of `man' from a vector of `king', and add a vector of `woman', then you will get a vector close to a vector of `queen'.

i) (4 points) Briefly describe a situation when one-hot encoding transformation is more desirable than distributed word vector transformation and explain the Reason.

- One hot encoding is a simple method to vectorize words, and is efficient enough to perform basic NLP tasks. On the other hand lots of data and training is required to create Distributed vectors, but it is much more efficient and is used for handling complex problems.
- For simple NLP tasks like Text Summarization, One-hot encoding is suitable to calculate term or work frequency vectors and represent them as matrices. For Text Summarization calculating the frequency of occurrence of words is made easy by vectorizing the words using one-hot encoding. So for smaller tasks like this which do not suffer from the curse of dimensionality, one-hot encoding is desirable.

ii) (4 points) Briefly describe a situation when distributed word vector transformation is more desirable than one-hot encoding transformation and explain the reason.

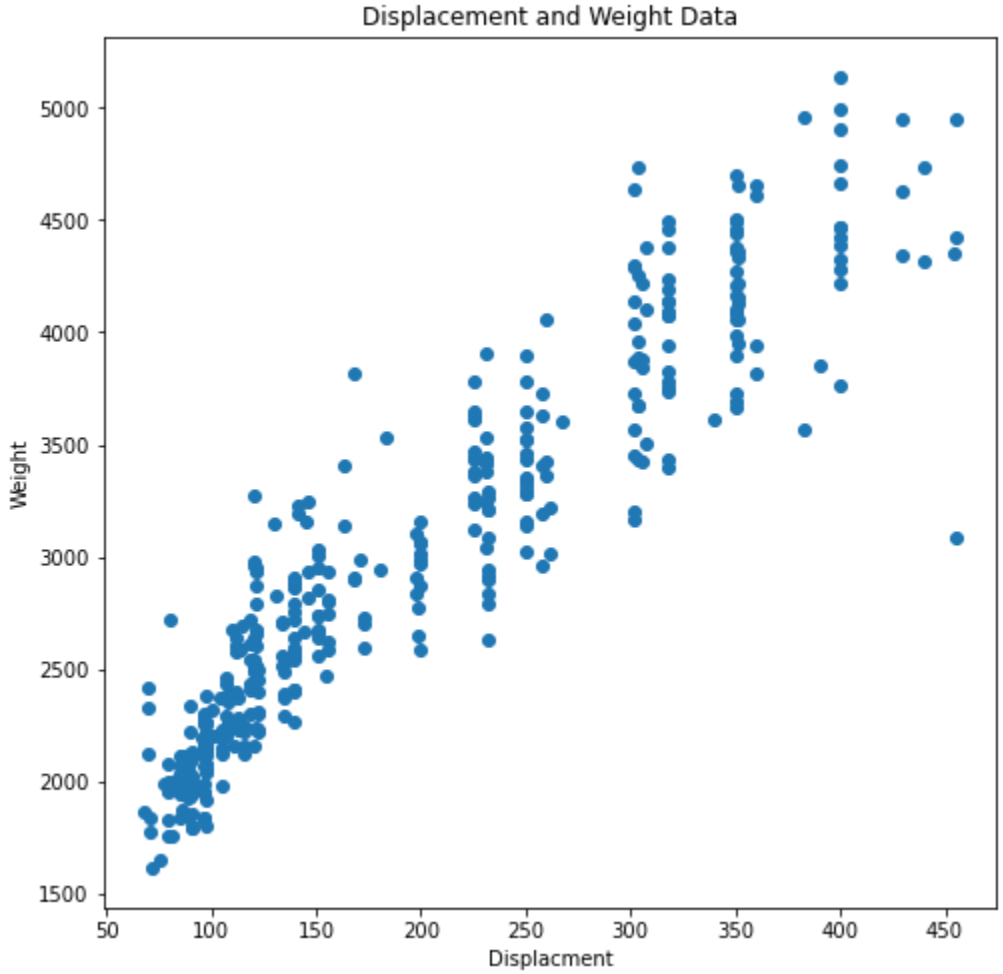
- A distributed approach to forming word vectors is much more efficient than one-hot encoding, especially at high volumes of data. With a one-hot encoder, adding much more data would exponentially increase the dimensionality of data and often lead to sparsity.
- Additionally, distributed word vectors hold the meaning of the words they represent in such a way that they can handle analogy based tasks. Tasks such as: *A* is to *B*, then *C* is to ?, since semantically similar words would be located closer to each other in vector space. So they are desirable in translation based tasks. Example, in translation between english and french: ‘Hello’ and ‘Bonjour’ would be closer to each other in the

dimensional space. In this application one-hot encoding can not be used as its vectors do not hold any meaning.

---

7) (30 points) [Distance] [John Wesley Hostetter] For this exercise, use the provided les “datanauto-mpg.data" and “datanauto-mpg.names" , which contains a list of 398 data instances. There are 9 attributes, including the class attribute; please refer to the included link for documentation. For this exercise, we will only be concerned with a select few { namely, the displacement and weight attributes. Write code in Python to perform the following tasks, please report your output and relevant code in the document le, and also include your code le (ends with .py) in the .zip le.

(a) (4 points) Data is not always readily available in .csv les, and sometimes must be appropriately formatted to facilitate data manipulation or analysis. Therefore, parse the .data file into a more accessible representation e.g. a Pandas DataFrame. You may consider beginning by examining the provided .data le for useful patterns to use with a regular expression. Then, generate a plot between the displacement and the weight of the observations. Label the axes (displacement should be x-axis and weight should be y-axis). Call this plot “Displacement and Weight Data". What general interpretation can you make from this plot?



From the above plot, we can infer that there is a positive correlation between the displacement and weight. This means that an increase in weight causes a likewise increase in displacement (and vice versa).

(b) (2 points) Compute the mean of the attributes displacement and weight. Define a data point called P such that P = (mean(displacement); mean(weight)).

```
[193.42587939698493, 2970.424623115578]
```

(c) (10 points) Compute the distance between P and the 398 data points using the following distance measures: 1) Euclidean distance, 2) Manhattan block metric, 3) Minkowski metric (for power=7), 4) Chebyshev distance, and 5) Cosine distance. List the closest 6 points for each distance.

Closest 6 points for euclidean distance are
{'distance': array([8.52323868]), 'coords': (200.0, 2965.0)}
{'distance': array([20.64980489]), 'coords': (200.0, 2990.0)}
{'distance': array([26.21470817]), 'coords': (171.0, 2984.0)}
{'distance': array([28.29865614]), 'coords': (181.0, 2945.0)}
{'distance': array([42.09193539]), 'coords': (200.0, 3012.0)}
{'distance': array([46.19928832]), 'coords': (232.0, 2945.0)}
Closest 6 points for cityblock distance are
{'distance': array([11.99874372]), 'coords': (200.0, 2965.0)}
{'distance': array([26.14949749]), 'coords': (200.0, 2990.0)}
{'distance': array([36.00125628]), 'coords': (171.0, 2984.0)}
{'distance': array([37.85050251]), 'coords': (181.0, 2945.0)}
{'distance': array([48.14949749]), 'coords': (200.0, 3012.0)}
{'distance': array([62.85050251]), 'coords': (151.0, 2950.0)}

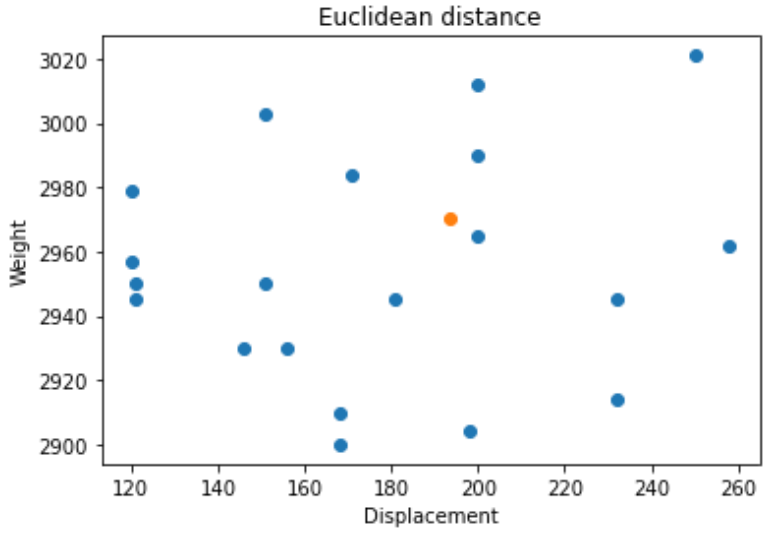
Closest 6 points for minkowski distance are
{'distance': array([8.52323868]), 'coords': (200.0, 2965.0)}
{'distance': array([20.64980489]), 'coords': (200.0, 2990.0)}
{'distance': array([26.21470817]), 'coords': (171.0, 2984.0)}
{'distance': array([28.29865614]), 'coords': (181.0, 2945.0)}
{'distance': array([42.09193539]), 'coords': (200.0, 3012.0)}
{'distance': array([46.19928832]), 'coords': (232.0, 2945.0)}

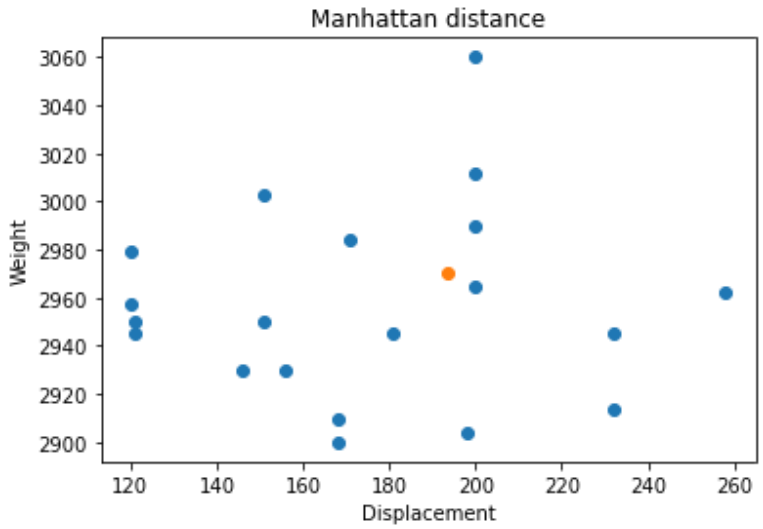
Closest 6 points for chebyshev distance are
{'distance': array([6.5741206]), 'coords': (200.0, 2965.0)}
{'distance': array([19.57537688]), 'coords': (200.0, 2990.0)}
{'distance': array([22.4258794]), 'coords': (171.0, 2984.0)}
{'distance': array([25.42462312]), 'coords': (181.0, 2945.0)}
{'distance': array([38.5741206]), 'coords': (232.0, 2945.0)}
{'distance': array([40.42462312]), 'coords': (156.0, 2930.0)}

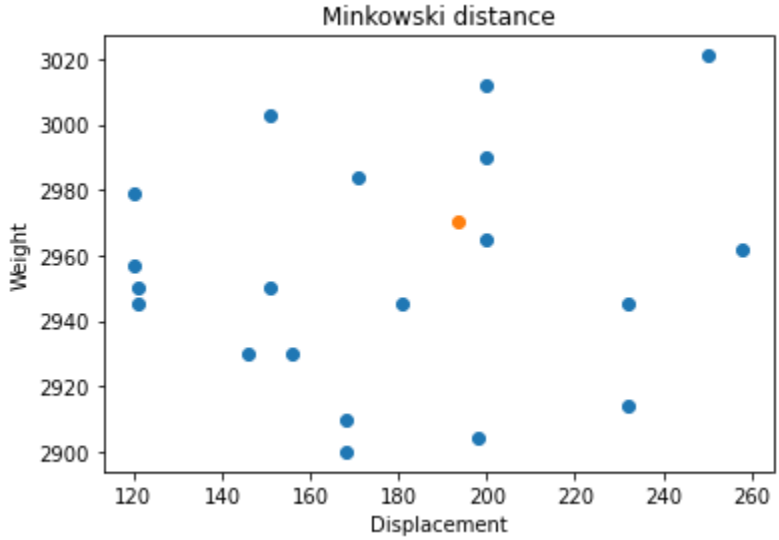
Closest 6 points for cosine distance are
{'distance': array([4.38926673e-12]), 'coords': (302.0, 4638.0)}
{'distance': array([4.26525704e-10]), 'coords': (200.0, 3070.0)}
{'distance': array([1.64560311e-08]), 'coords': (225.0, 3465.0)}
{'distance': array([2.60637032e-08]), 'coords': (231.0, 3535.0)}
{'distance': array([2.90891299e-08]), 'coords': (200.0, 3060.0)}
{'distance': array([4.72588364e-08]), 'coords': (225.0, 3439.0)

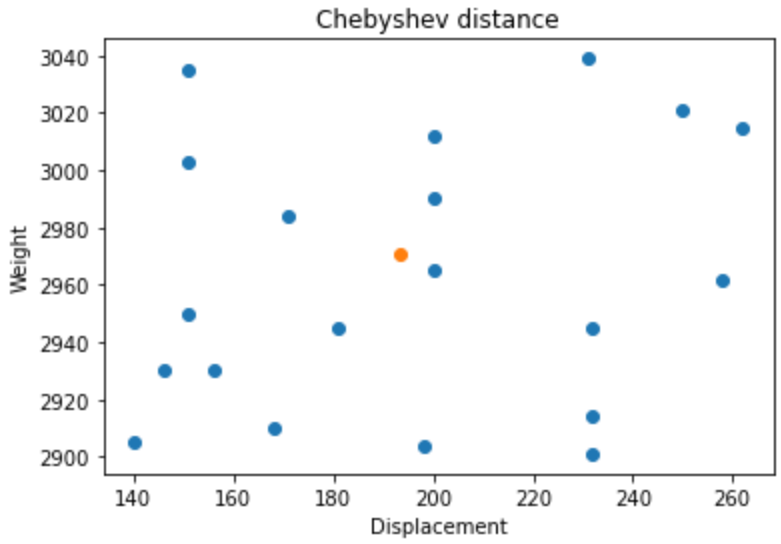
(d) For each distance measure, identify the 20 points from the dataset that are the closest to the point P from (b). (You are allowed to use any package functions to calculate the distances.)

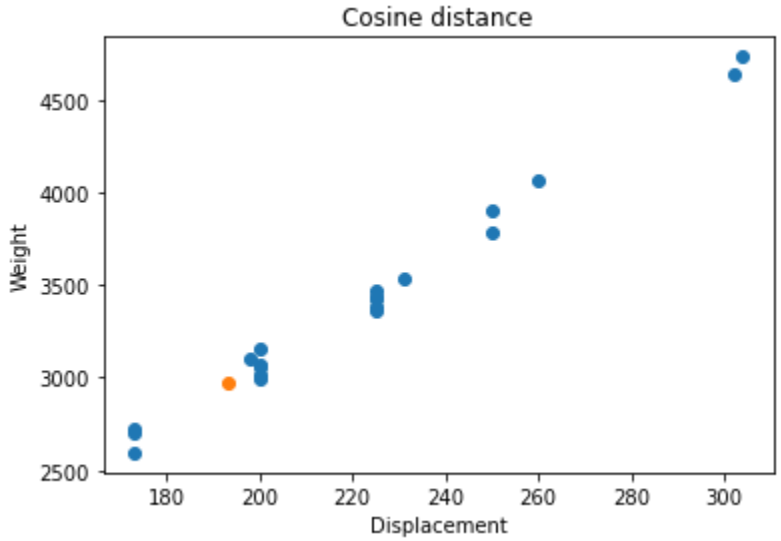
i. (10 points) Create plots, one for each distance measure. Place P on the plot and mark the 20 closest points. To mark them, you could use different colors or shapes. Make sure the points can be uniquely identified




Manhattan distance
Weight
Displacement
3060
3040
3020
3000
2980
2960
2940
2920
2900
120
140
160
180
200
220
240
260


Minkowski distance
Weight
Displacement
3020
3000
2980
2960
2940
2920
2900
120
140
160
180
200
220
240
260





ii. (4 points) Verify if the set of points is the same across all the distance measures. If there is any big difference, briefly explain why it is.

We can infer that the nearest 20 points using cosine distance varies significantly from the rest of the distance measures.The Euclidean distance determines the shortest straight line distance between the two points, however the Cosine Distance determines the angular distance of a line drawn from the origin to the point in consideration with the coordinate axes. The points are then grouped together based on closeness of angle and not conventional straight line distance.The distance measured is the cos inverse of the dot product.
