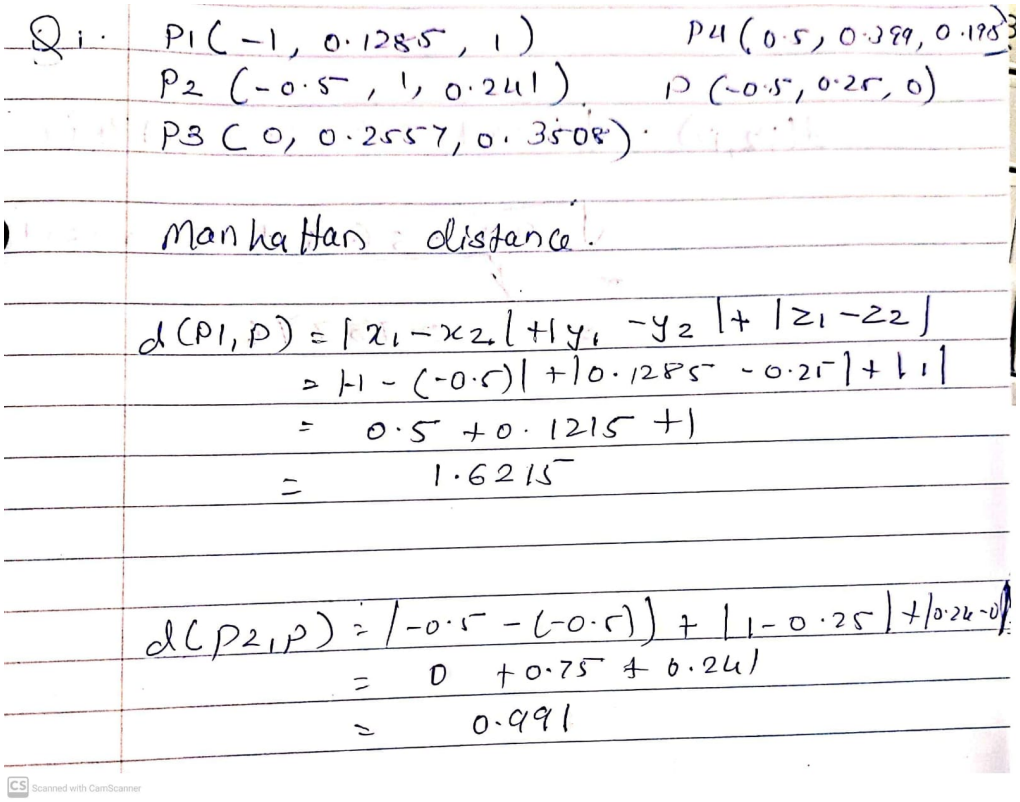
CSCI 552 (Spring 2021)

Homework #1

Name – Khushboo Mantri
Class – Data visualization
Handout: Thursday, Feb. 11, 2021
Due: 11:59 pm, Thursday, Feb. 25, 2021

1. Given a set of 3D data points: $P_1 = (-1, 0.1285, 1)$, $P_2 = (-0.5, 1, 0.241)$, $P_3 = (0, 0.2557, 0.3508)$, $P_4 = (0.5, 0.399, 0.198)$, compute the Manhattan distances, Euclidean distance, and Cosine distances between each of these points to a point $P = (-0.5, 0.25, 0)$.



Q i. P1 (−1, 0.1285, 1) P4 (0.5, 0.399, 0.198)
P2 (−0.5, 1, 0.241) P (−0.5, 0.25, 0)
P3 (0, 0.2557, 0.3508)

Manhattan distance.

$d(P_1, P) = |x_1 - x_2| + |y_1 - y_2| + |z_1 - z_2|$
$= |-1 - (-0.5)| + |0.1285 - 0.25| + |1|$
$= 0.5 + 0.1215 + 1$
$= 1.6215$

$d(P_2, P) = |-0.5 - (-0.5)| + |1 - 0.25| + |0.24 - 0|$
$= 0 + 0.75 + 0.241$
$= 0.991$

Scanned with CamScanner

$d(P_3, P) = |0+0.5| + |0.2557 - 0.25| + |0.3508 - 0|$

$= 0.5 + 0.0057 + 0.3508$

$= 0.8565$

$d(P_4, P) = |0.5+0.5| + |0.399 - 0.25| + |0.198 - 0|$

$= 1 + 0.149 + 0.198$

$= 1.347$

Euclidean distance.

$d(P_1, P) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$

$= \sqrt{(-1 - (-0.5))^2 + (0.1265 - 0.25)^2 + (1-0)^2}$

$= \sqrt{0.25 + 0.01476225 + 1}$

$= 1.124$

$d(P_2, P) = \sqrt{(-0.5+0.5)^2 + (1-0.25)^2 + (0.241-0)^2}$

$= \sqrt{0 + 0.5625 + 0.058081}$

$= 0.78776$

$d(P_3, P) = \sqrt{(0+0.5)^2 + (0.2557-0.25)^2 + (0.3508-0)^2}$

$= \sqrt{0.25 + 0.00003249 + 0.12386064}$

$= 0.610813498$

$d(P_4, P) = \sqrt{(0.5+0.5)^2 + (0.399-0.25)^2 + (0.198-0)^2}$

$= \sqrt{1 + 0.022201 + 0.039204}$

$= 1.030245116$

cosine distance

$$d(P_1, P) = \frac{x_1 \cdot x_2 + y_1 \cdot y_2 + z_1 \cdot z_2}{\sqrt{x_1^2 + y_1^2 + z_1^2} \times \sqrt{x_2^2 + y_2^2 + z_2^2}}$$

$$= \frac{(-1)(-0.5) + (0.1285)(0.25) + (1)(0)}{\sqrt{(-1)^2 + (0.1285)^2 + (1)^2} \times \sqrt{(-0.5)^2 + (0.25)^2 + 0}}$$

$$= \frac{0.5 + 0.032125 + 0}{(1.420039)(0.55901)}$$

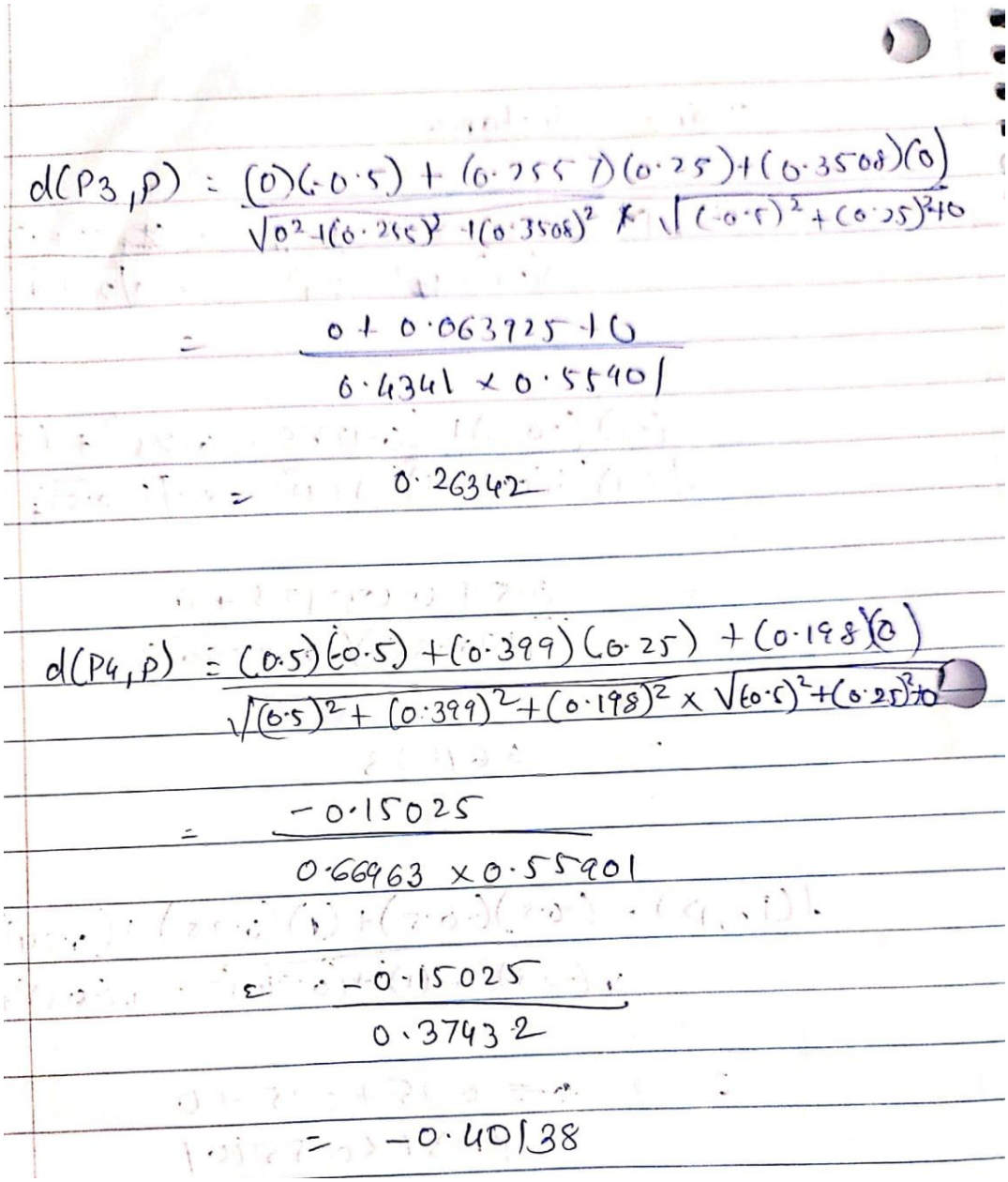
$$= 0.67033$$

$$d(P_2, P) = \frac{(-0.5)(-0.5) + (1)(0.25) + (0.241)(0)}{\sqrt{(-0.5)^2 + (1)^2 + (0.241)^2} \times \sqrt{(-0.5)^2 + (0.25)^2 + 0}}$$

$$= \frac{0.25 + 0.25 + 0}{1.1487 \times 0.55901}$$

$$= \frac{0.5}{0.639339737}$$

$$= 0.78205681$$



2. Prove that the distance formula for categorical data, $d(i, j) = \frac{p-m}{p}$, satisfies the triangle inequality property.

Triangle inequality

$d(i,j) = \frac{p-m}{p}$ p, total types
m, same types

consider a triangle ΔABC.

where $\frac{p-m}{p} = \frac{(A \cup B)}{(A \cap B)}$

When $\frac{(A \cup B)}{(A \cap B)}$ is called Symmetric differ of ΔABC.

$\therefore d(A,B) = |A \Delta B|$

$(A \Delta B) \Delta (B \Delta C) = A \Delta C$; ← Symmetric diff

$\therefore A \Delta C \subseteq (A \Delta B) \cup (B \Delta C)$

$\therefore d(A,C) \leq d(A,B) + d(B,C)$

3. The weight (kg) and height (cm) measurements for a group of students are: (45, 149), (48, 153), (47, 156), (49, 156), (53, 161), (52, 162), (56, 162), (49, 162), (50, 164), (52, 165).

(a) Compute the following height statistics: mean, median, standard deviation, the first quartile, and the third quartile.

Q3: (45, 149)
(48, 153)
(47, 156)
(49, 156)
(53, 161)
(52, 162)
(56, 162)
(49, 162)
(50, 164)
(52, 165)

1) Mean of Height

$\frac{149 + 153 + 156 + 156 + 161 + 162 + 162 + 162 + 164 + 165}{10}$

$= \frac{1590}{10}$

$= 159$ cm

1.2 median of height

$\frac{161 + 162}{2} = \frac{323}{2} = 161.5$ cm

Standard deviation

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$(x_i - \bar{x})$	10	6	3	3	2	3	3	5	6
$(x_i - \bar{x})^2$	100	36	9	9	4	9	9	25	36

$$\therefore \sum (x_i - \bar{x})^2 = 246.$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$= \frac{1}{10}(246)$$

$$= \frac{246}{10}$$

$$= 24.6$$

$$\sigma = \sqrt{24.6}$$

$$\sigma = 4.96$$

(b) Manually draw a box plot for the height data, and a scatter plot on a weight-height plane.

1.4 First quartile : $\frac{n}{4}$ $Q1 = X_{int\,qz}\left(\frac{n+1}{4}\right) +$

$X_{int\,qz}\left(\frac{n+1}{4}\right) + 1$

$Q1 = 155.25$ cm

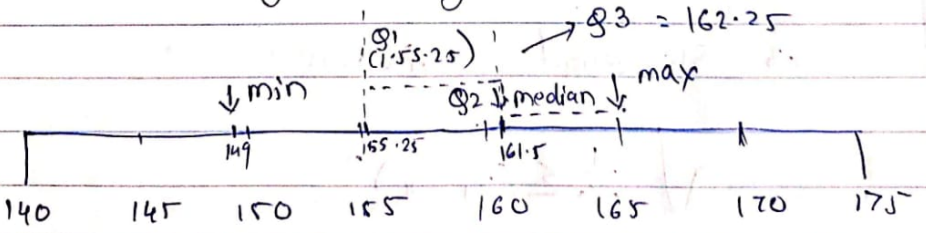
1.5 Third quartile : formula $Q3\ X_{int\,qz}\left(\frac{3(n+1)}{4}\right)$

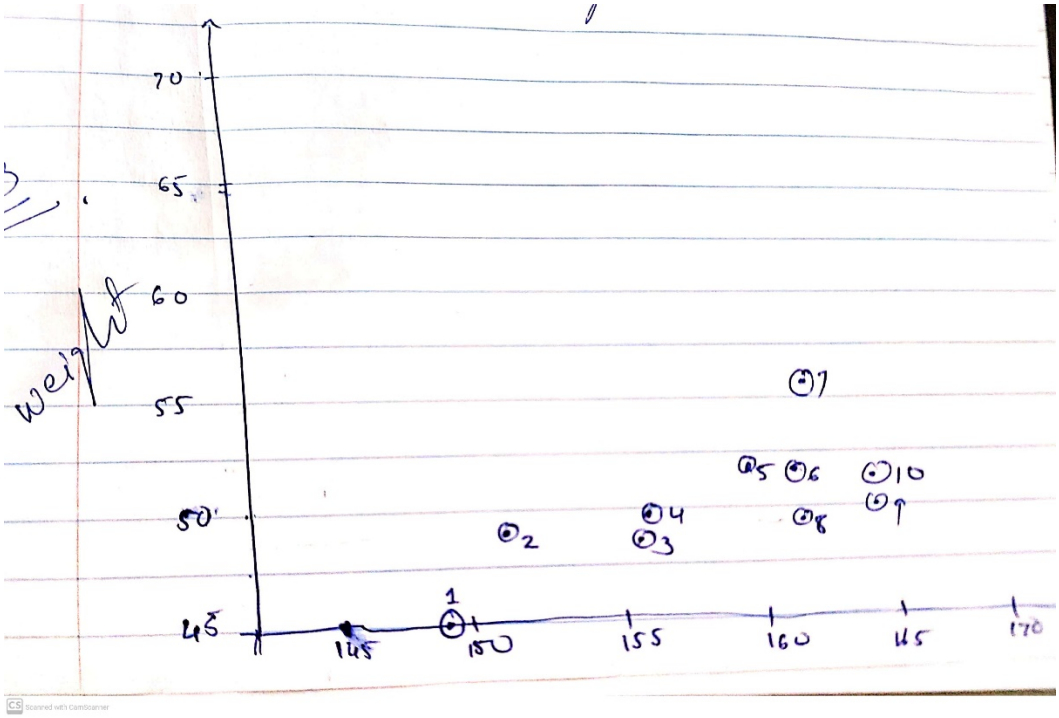
$+ X_{int\,qz}\left(\frac{3(n+1)}{4}\right) +$

$X_{int\,qz}\left(\frac{3(n+1)}{4}\right)$

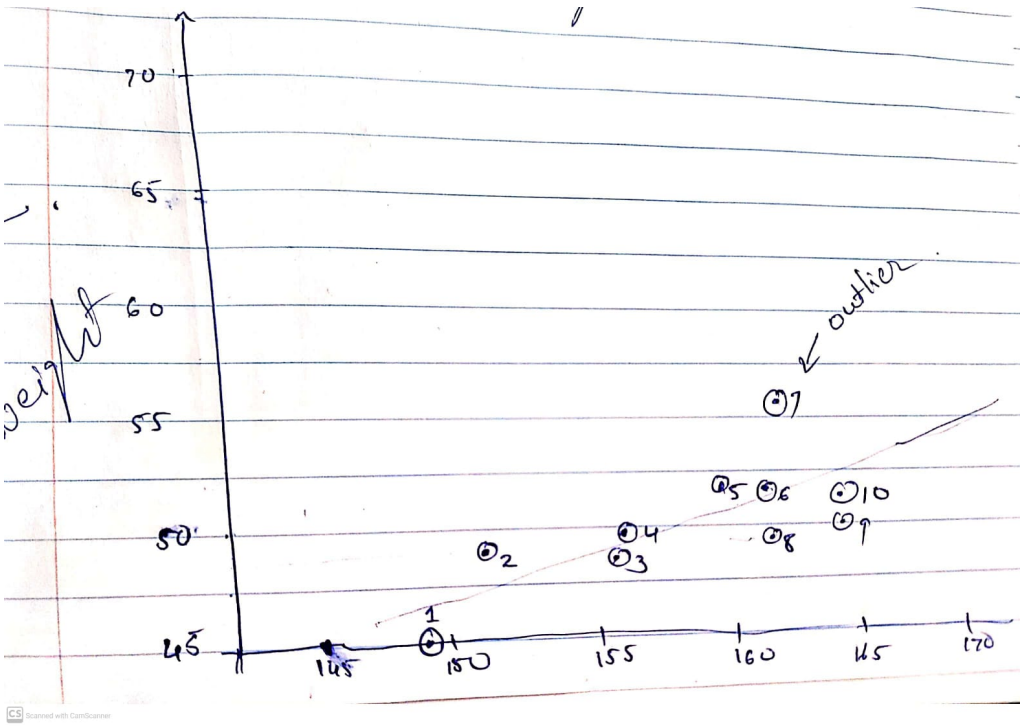
$= 162.5$ cm

2) Box plot for height data





(c) Manually draw a linear regression curve, and indicate if there are outliers.



4. Perform an Internet search to find three interesting repositories of publicly available data. Study these data sources and write one or two paragraphs for each repository describing the nature, characteristics and structure of the data. For example, how was the data set collected? what types of attributes and features are there in the data set? Does it have spatial or temporal attributes? What file format is used? etc.

Iris Dataset - It contains these columns: SepalLength, SepalWidth, PetalLength, PetalWidth, Name. The data set consists of 50 samples from each of three species of Iris (Iris Setosa, Iris virginica, and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. The file is in csv format. Link of dataset - https://github.com/rashida048/Datasets/blob/master/iris.csv

Airline – safety – This dataset was created with a purpose that Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past? The dataset contains following field with the definition of the field

Header	Definition
airline	Airline (asterisk indicates that regional subsidiaries are included)
avail_seat_km_per_week	Available seat kilometers flown every week
incidents_85_99	Total number of incidents, 1985–1999
fatal_accidents_85_99	Total number of fatal accidents, 1985–1999
fatalities_85_99	Total number of fatalities, 1985–1999
incidents_00_14	Total number of incidents, 2000–2014
fatal_accidents_00_14	Total number of fatal accidents, 2000–2014
fatalities_00_14	Total number of fatalities, 2000–2014

Link to dataset - https://github.com/fivethirtyeight/data/tree/master/airline-safety

Canada Immigration Dataset - This dataset provides information about how many immigrants came from which country by year. This file has total 196 rows and 51 attributes. Format of dataset is xlsx. Link for dataset - https://github.com/rashida048/Datasets/blob/master/Canada.xlsx

Canada Immigration Dataset - Dataset for how Canada sends out invitation to invite Immigrants The dataset provides the latest information available publicly on IRCC website. It tells what immigration programs are being targeted, how many people are being invited and what is the score cutoff for each rounds of invitation. A lower CRS Score means that more people will be happy, and immigrants are constantly on a lookout for a trend where cores would fall. Based on historic data, you can predict whether the scores are going up or down in 2021. The dataset contain 12 columns, Link to dataset - https://www.kaggle.com/umerkk12/canada-immigration-dataset.