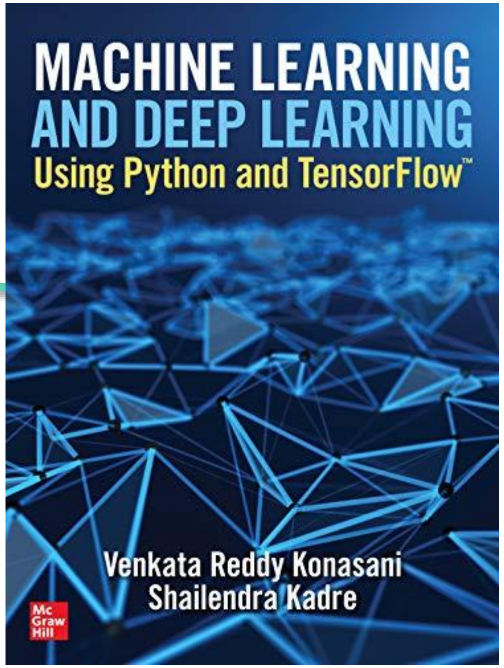
DV Analytics
Transforming You

Basic Statistics

Venkata Reddy Konasani

Chapter 2 in the book



MACHINE LEARNING AND DEEP LEARNING
Using Python and TensorFlow™

Venkata Reddy Konasani
Shailendra Kadre

Mc Graw Hill

Contents

- Descriptive statistics
  - Central Tendency
  - Variance
- Percentiles
- Quartiles
- Outlier Detection
- Box-plot

DV Analytics
Transforming You

Descriptive statistics

Descriptive statistics

- The basic descriptive statistics to give us an idea on the variables and their distributions
- Permit the analyst to describe many pieces of data with a few indices
- Central tendencies
  - Mean
  - Median
- Dispersion
  - Range
  - Variance
  - Standard deviation

DV Analytics
Transforming You

Central tendencies: Mean and Median

Central tendencies

- Mean
  - The arithmetic mean
  - Sum of values/ Count of values
  - Gives a quick idea on average of a variable

Mean in Python

Import “Census Income Data/Income_data.csv”

```
gain_mean=Income["capital-gain"].mean()
gain_mean
```

Guess the mean

1.5,1.7,1.9,0.8,0.8,1.2,1.9,1.4, 9 , 0.7 , 1.1

Median

- Mean is not a good measure in presence of outliers
- For example Consider below data vector
  - 1.5,1.7,1.9,0.8,0.8,1.2,1.9,1.4, 9 , 0.7 , 1.1
- 90% of the above values are less than 2, but the mean of above vector is 2
- There is an unusual value in the above data vector i.e 9
- It is also known as outlier.
- Mean is not the true middle value in presence of outliers. Mean is very much effected by the outliers.
- We use median, the true middle value in such cases
- Sort the data either in ascending or descending order

Median

1.5	0.7
1.7	0.8
1.9	0.8
0.8	1.1
0.8	1.2
1.2	**1.4**
1.9	1.5
1.4	1.7
9	1.9
0.7	1.9
1.1	9

- Mean of the data is 2
- Median of the data is 1.4
- Even if we have the outlier as 90, we will have the same median
- Median is a positional measure, it doesn't really depend on outliers
- When there are no outliers then mean and median will be nearly equal
- When mean is not equal to median it gives us an idea on presence of outliers in the data

Mean and Median

Import “Census Income Data/Income_data.csv”

```
#Mean and Median on python
gain_mean=Income["capital-gain"].mean()
gain_mean

gain_median=Income["capital-gain"].median()
gain_median
```

Mean is far away from median. Looks like there are outliers, we need to look at percentiles and box plot.

DV Analytics
Transforming You

Dispersion Measures : Variance and Standard Deviation

Dispersion

- Just knowing the central tendency is not enough.
- Two variables might have same mean, but they might be very different.
- Look at these two variables. Profit details of two companies A & B for last 14 Quarters in MMs

															Mean
Company A	43	44	0	25	20	35	-8	13	-10	-8	32	11	-8	21	15
Company B	17	15	12	17	15	18	12	15	12	13	18	18	14	14	15

- Though the average profit is 15 in both the cases
- Company B has performed consistently than company A.
- There was even loses for company A
- Measures of dispersion become very vital in such cases

Variance and Standard deviation

- Dispersion is the quantification of deviation of each point from the mean value.
- Variance is average of squared distances of each point from the mean
- Variance is a fairly good measure of dispersion.
- Variance in profit for company A is 352 and Company B is 4.9

Value	Value-Mean	(Value-Mean)^2
43	28	784
44	29	841
0	-15	225
25	10	100
20	5	25
35	20	400
-8	-23	529
13	-2	4
-10	-25	625
-8	-23	529
32	17	289
11	-4	16
-8	-23	529
21	6	36
15.0		**352**

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n}$$

Value	Value-Mean	(Value-Mean)^2
17	2	4
15	0	0
12	-3	9
17	2	4
15	0	0
18	3	9
12	-3	9
15	0	0
12	-3	9
13	-2	4
18	3	9
18	3	9
14	-1	1
14	-1	1
15.0		**4.9**

Standard Deviation

- Standard deviation is just the square root of variance
- Variance gives a good idea on dispersion, but it is of the order of squares.
- Its very clear from the formula, variance unites are squared than that of original data.
- Standard deviation is the variance measure that is in the same units as the original data

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$

LAB: Variance and Standard deviation

- Dataset: "./Online Retail Sales Data/Online Retail.csv”
- What is the variance and s.d of “UnitPrice”
- What is the variance and s.d of “Quantity”

```
Online_Retail=pd.read_csv("D:\\Datasets\\Online_Retail_Sales_Data\\Online
Retail.csv", encoding = "ISO-8859-1")
```

LAB: Variance and Standard deviation

```
#var and sd UnitPrice
Online_Retail_germany=Online_Retail[Online_Retail['Country']=='Germany']
Online_Retail_france=Online_Retail[Online_Retail['Country']=='France']

var_UnitPrice_germany=Online_Retail_germany['UnitPrice'].var()
print("Variance of UnitPrice Germany", var_UnitPrice_germany)

var_UnitPrice_france=Online_Retail_france['UnitPrice'].var()
print("Variance of UnitPrice France", var_UnitPrice_france)
```

Percentiles & Quartiles

Percentiles

- A student attended an exam along with 1000 others.
  - He got 68% marks? How good or bad he performed in the exam?
  - What will be his rank overall?
  - What will be his rank if there were 100 students overall?
- For example, with 68 marks, he stood at 90th position. There are 910 students who got less than 68, only 89 students got more marks than him
- He is standing at 91 percentile.
- Instead of stating 68 marks, 91% gives a good idea on his performance
- Percentiles make the data easy to read

Percentiles

- pth percentile: p percent of observations below it, (100 - p)% above it.
- Marks are 40 but percentile is 80%, what does this mean?
- 80% of CAT exam percentile means
  - 20% are above & 80% are below
- Percentiles help us in getting an idea on outliers.
- For example the highest income value is 400,000 but 95th percentile is 20,000 only. That means 95% of the values are less than 20,000. So the values near 400,000 are clearly outliers

Percentiles

```
Income['capital-gain'].quantile([0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1])
```

Lab: Outlier detection

- https://www.kaggle.com/c/GiveMeSomeCredit
- Import “Give me some Credit\cs-training.csv”
- Look at the percentiles of the variable monthly_utilization
- Are there any outliers?

Code: Outlier detection

```
loans['monthly_utilization'].quantile([0, 0.1, 0.2, 0.3,
0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1])
```

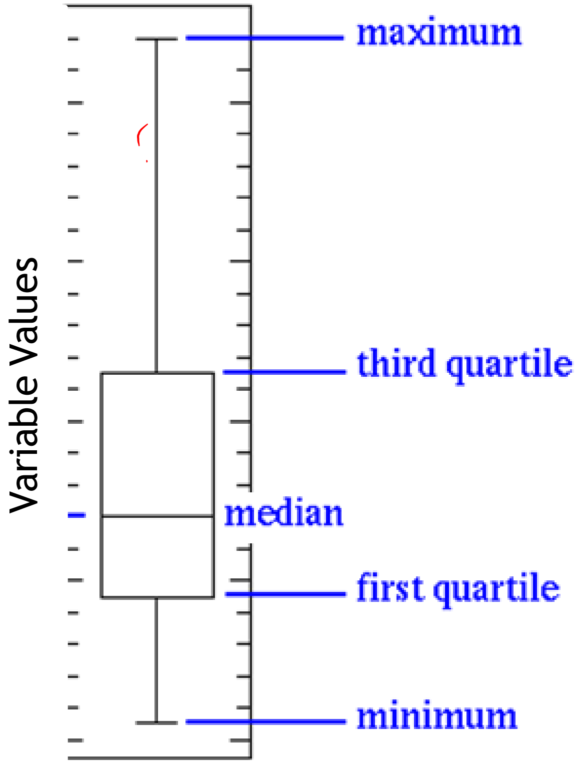
Quartiles

- Percentiles divide the whole population into 100 groups where as quartiles divide the population into 4 groups
- p = 25: First Quartile or Lower quartile (LQ)
- p = 50: second quartile or Median
- p = 75: Third Quartile or Upper quartile (UQ)

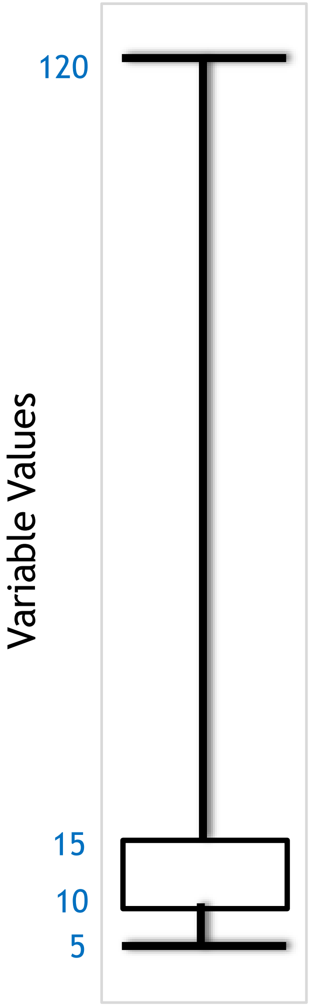
Box plots and outlier detection

Box plots and outlier detection

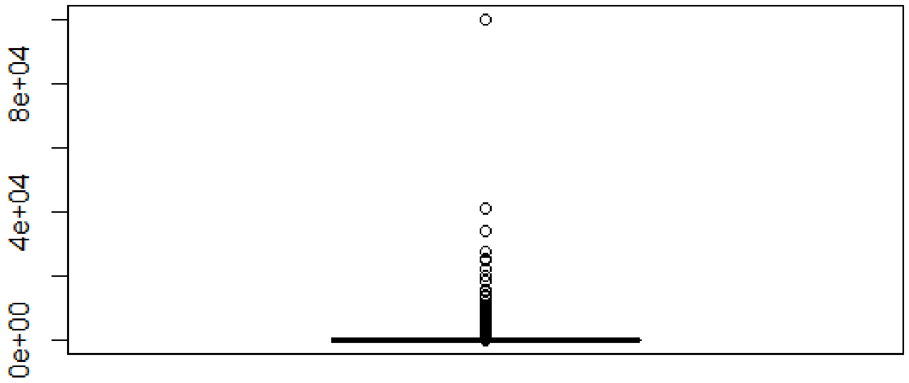
- Box plots have box from LQ to UQ, with median marked.
- They portray a five-number graphical summary of the data Minimum, LQ, Median, UQ, Maximum
- Helps us to get an idea on the data distribution
- Helps us to identify the outliers easily
- 25% of the population is below first quartile,
- 75% of the population is below third quartile
- If the box is pushed to one side and some values are far away from the box then it’s a clear indication of outliers



Box plots and outlier detection



- Some set of values far away from box, is gives us a clear indication of outliers.
- In this example the minimum is 5, maximum is 120, and 75% of the values are less than 15
- Still there are some records reaching 120. Hence a clear indication of outliers
- Sometimes the outliers are so evident that, the box appear to be a horizontal line in box plot.



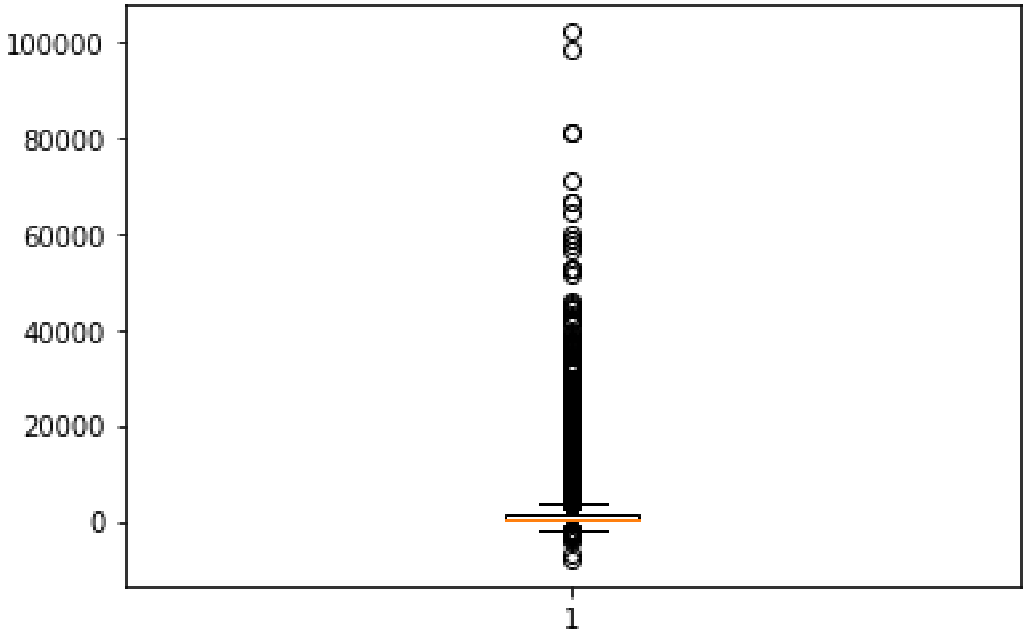
Box plots and outlier detection

Dataset: “./Bank Tele Marketing/bank_market.csv”

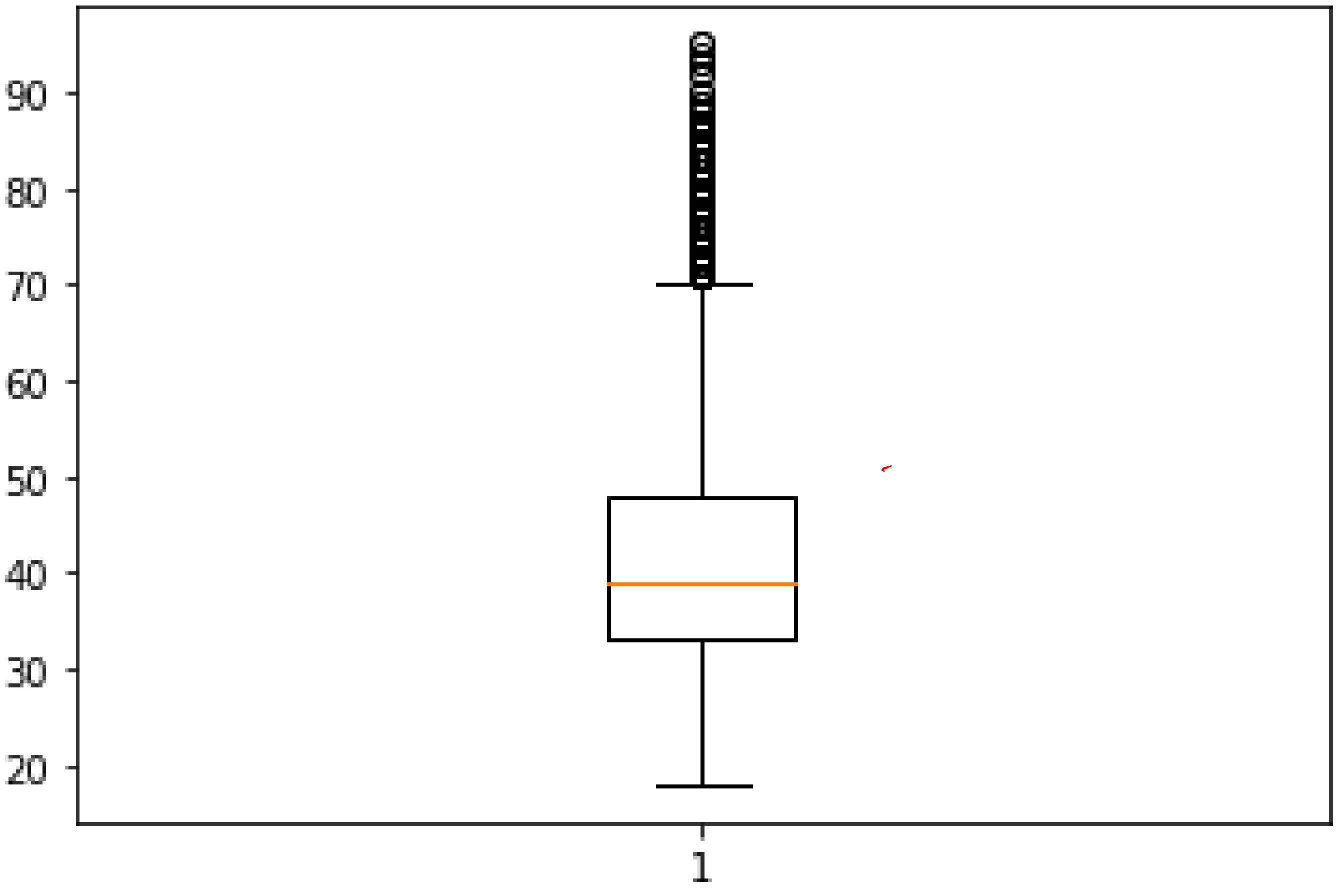
```
import matplotlib.pyplot as plt
plt.boxplot(bank.balance)
```

```
plt.boxplot(bank.age)
```

Output - Balance



Output - Age



LAB: Box plots and outlier detection

- Dataset: “./Bank Marketing/bank_market.csv”
- Draw a box plot for balance variable
- Do you suspect any outliers in balance ?
- Get relevant percentiles and see their distribution.
- Draw a box plot for age variable
- Do you suspect any outliers in age?
- Get relevant percentiles and see their distribution.

Conclusion

- In this session we discussed some basic data reporting
- Studying descriptive statistics is essential before we start our advanced modeling. It gives us an idea on variable distribution