**Assignment 2 – Basic Statistics in Python**

**NOTE:**

1. **Solve all the questions in a Jupyter Notebook**
2. **The answers should be properly annotated**
3. Given below is the data of 20 patients (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6813708/>)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sr. No.** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** |
| Gender | M | M | M | M | M | M | M | M | M | M | F | F | F | F | F | F | F | F | F | F |
| Age groups | 1 | 2 | 3 | 1 | 1 | 3 | 3 | 2 | 1 | 3 | 3 | 2 | 2 | 2 | 3 | 1 | 1 | 3 | 3 | 2 |
| BMI | 23 | 25 | 26 | 24 | 23 | 27 | 27 | 24 | 21 | 28 | 26 | 26 | 23 | 24 | 25 | 22 | 19 | 25 | 26 | 25 |
| DBP (B/L) | 68 | 80 | 85 | 75 | 77 | 82 | 80 | 82 | 80 | 74 | 75 | 86 | 85 | 87 | 83 | 79 | 73 | 77 | 82 | 81 |
| DBP (at 30 min) | 75 | 79 | 90 | 80 | 81 | 89 | 92 | 90 | 80 | 80 | 89 | 88 | 87 | 91 | 88 | 78 | 74 | 84 | 83 | 80 |
| DBP (at 60 min) | 70 | 78 | 85 | 75 | 72 | 72 | 75 | 85 | 80 | 75 | 85 | 80 | 82 | 82 | 85 | 74 | 87 | 86 | 85 | 72 |

Age groups: 1 (<30 years), 2 (30-50 years), 3 (>50 years).

Gender: M=Male, F=Female,

BMI=Body mass index,

DBP=Diastolic blood pressure,

B/L=Baseline, min=Minute

1. Represent the above data in a Pandas data frame with 20 observations and 6 attributes
2. Identify the data in terms of categorical, numeric etc.
3. Write descriptive statistics for the numeric data
4. Analyse and represent the categorical variables in terms of percentage through a pie chart
5. Hypothetical data representing the measurements of bonding strength according to three different types of resin is given below:

|  |  |  |
| --- | --- | --- |
| A | B | C |
| 19.7  20.1  21.3  23.5  9.3  27.1  11.6  12.2  15.9  17.0  17.2  18.4  19.4  23.4  2.0 | 23.0  24.5  24.6  27.1  12.0  27.8  12.8  16.2  19.8  22.4  23.6  25.3  27.9  4.6  35.2 | 21.6  25.5  25.9  30.7  3.0  16.5  22.7  24.2  26.2  28.4  28.5  30.7  32.2  33.8  34.5 |

Using ANOVA, identify if there is any significant difference between the mean strength of the three groups.

1. The table below lists the annual income of Kirana store owners in New-Delhi as told by them. We want to estimate the average annual income of a Kirana store owner in New-Delhi.

What would be your estimate of the average annual income of a Kirana Store Owner in Delhi?

Construct a 99%, 95% and 90% confidence interval for the average annual income of a Kirana store owner in New Delhi.

|  |  |  |
| --- | --- | --- |
| **Kirana Store Owner** | |  |
| **S.No** | **Name** | **Annual Income** |
| 1 | RAHUL | 802000 |
| 2 | ISHIRA | 731000 |
| 3 | DARSHAN | 963000 |
| 4 | DHANANJAY | 832000 |
| 5 | RUSHABH | 979000 |
| 6 | PARTH | 498000 |
| 7 | YOUNUS | 805000 |
| 8 | RIDDHI | 531000 |
| 9 | TAVISHI | 462000 |
| 10 | AKHIL | 744000 |
| 11 | MOHNISH | 790000 |
| 12 | AJAY | 746000 |
| 13 | SIMRAN | 486000 |
| 14 | VANSHIKA | 890000 |
| 15 | VISHAL | 755000 |
| 16 | UTKARSH | 822000 |
| 17 | DESHPANDE | 804000 |
| 18 | PANKAJ | 883000 |
| 19 | MOHIT | 393000 |
| 20 | KATTA | 748000 |
| 21 | PALWINDER | 957000 |
| 22 | PALLAVI | 817000 |
| 23 | ASTHA | 618000 |
| 24 | HARSHVARDHAN | 940000 |
| 25 | VANI | 393000 |
| 26 | YOGESH | 604000 |
| 27 | RAJAT | 622000 |
| 28 | ANAMIKA | 710000 |
| 29 | PRAGYAN | 620000 |
| 30 | MAHENDAR | 765000 |
| 31 | BHASKAR | 348000 |
| 32 | GANESH | 399000 |
| 33 | RASHI | 316000 |
| 34 | AASHRAYA | 874000 |
| 35 | DHEERAJ | 395000 |
| 36 | TEENA | 399000 |
| 37 | ARUN | 943000 |
| 38 | KHUSHBU | 949000 |
| 39 | TRIPTI | 732000 |
| 40 | ROHIT | 634000 |
| 41 | SOUDIP | 727000 |
| 42 | DHRUV | 704000 |
| 43 | HARSH | 341000 |
| 44 | RAHUL | 980000 |
| 45 | VEENA | 935000 |
| 46 | VIPIN | 377000 |
| 47 | RAHUL | 877000 |
| 48 | RADHIKA | 840000 |
| 49 | AJINKYA | 846000 |
| 50 | JISHAN | 350000 |

1. There was a heated argument between two groups of people about the proportion of people riding 2-wheelers without wearing helmet. One group claimed that traffic laws are enforced more in larger cities like Ahmedabad than in relatively smaller cities like Anand. The onus of proof lies with the group which is making the claim, and hence the claim would go in the alternate hypothesis.

(a) 300 2-wheelers were randomly sampled in Anand, 60 among those were found to be wearing helmets. 400 2-wheelers were randomly sampled in Ahmedabad, 90 among those were found to be wearing helmets. Conduct a hypothesis test of equality of proportions, and give the verdict at significance level 1%, 5% and 10%.

(b) Would your answer to the above part change if the claim (alternative hypothesis) was that traffic laws are enforced differently for Ahmedabad and Anand. Write the null and alternate hypothesis, and give the verdict at significance level 1%, 5% and 10% by calculating the p-value.