1st Assignment:

Please start your Assignemnt 1 test, when you will find an answer for all questions.

You will have only ONE attempt.

You will have 3 hours for drag-and-dropping your answer into Moodle.

The deadline is 5th January 2025 23:59. No extensions are allowed

All the questions highlighted in blue are the questions that you will have in the 1st assignment test.

A) Import relevant libraries:

import pandas as pd

import numpy as np

import matplotlib as mpl

import seaborn as sns

import scipy as sp

ds = pd.read\_excel('YOUR PATH')

ds

1) How many rows and columns do you have in your initial dataset? ? and ? 20 and 10908

B) Remove duplicates in the data. Be careful when selecting duplicates - you need to remove complete duplicates. If the IDs match, then this is not considered as duplicate.

2) How many rows do you have after step B? ?

C) Fill in the empty fields (Missing data) in certain columns with values ​​(Imputation) - in some cases, you need to think through the logic.

3) How many missing values do you have in the following columns before any fillings? Hint: You can find an answer, if you will do Step C.

|  |  |
| --- | --- |
| [Город] City | ? |
| [Льготы] Benefits | ? |
| [Регион\_ДТП] Region\_Accident | ? |

Hint 2: Please fill missing collumns with the following logic by using ‘fillna’ method:

'Льготы': 'Нет льгот',

    'Дата\_проишествия': 'Нет проишествия',

    'Регион\_ДТП': 'Нет ДТП',

    'Сумма\_убытка': 0

D) Focus your project only on passenger cars no older than 25 years. Exclude buses, trailers and motorcycles. \* Keep in mind that the dataset is from 2013.

4) Now, how many rows do you have after step D? ?

E) Tidy up the values ​​in the [City] Город and [Year of manufacture] Год выпуска columns. For the [year of manufacture], information about the [Year] is enough, without the day and month. Hint: for example, you can write in the following way:

# Convert cities to lowercase and remove extra spaces

sorted\_df['Город'] = sorted\_df['Город'].str.strip().str.lower()

F) Exclude from the calculations insurance records for less than 30 days - for them, the insurance policy is formed differently. Make sure that when calculating the model, the price is correct depending on the insurance time. Hint: for this step, please do the followings:

* Convert start and end date to datetime format
* Calculating the duration of insurance in days. For example:

sorted\_df['Длительность страхования'] = (sorted\_df['Конец страхования'] - sorted\_df['Начало страхования']).dt.days + 1;

* Filter records, excluding those where the insurance duration is less than 30 days

sorted\_df = sorted\_df[sorted\_df['Длительность страхования'] >= 30]

5) How many rows do you have after step F? ?

G) For each field, perform basic statistics (e.g. mean, quartiles, standard deviation) - draw initial conclusions based on the results.

6) If you successfully implemented step G, please fulfill below the initial conclusion:

It is obvious that 25% (of first quartile) of [Возраст] Age is ?. The average mean of [Стаж Вождения] Driving Experience is ? (rounded up 2 digits after floating point). The Maximum [Страховая Премия] Insurance Premium is ?. Standard deviation of [КБМ] CBM is ? (rounded up 2 digits after floating point).

H) Perform basic analytics, answer the following questions:

H.1 How many cars does one person have at most

7) How many cars does one person have at most: ?

H.2 Find the ratio of accidents by gender - Male and Female

8) Male: ?%

Female: ?%

Please round up 2 digits after floating point.

H.3 How many cars were insured in the largest cities of Kazakhstan - Astana (in the Nur-Sultan dataset), Almaty, Shymkent and Atyrau

9) Number of cars that purchased insurance in major cities: ?

Hint: you need to sum up 4 Major cities listed above.

H.4 What are the 3 most popular car models among Kazakhstanis?

10) What are the 3 most popular car models among Kazakhstanis? (list please in descending order):

|  |  |  |
| --- | --- | --- |
| № | Car Model | Count |
| 1 | ? | ? |
| 2 | ? | ? |
| 3 | ? | ? |

H.5 For what period do Kazakhstanis usually buy insurance?

11) Top 5 most common insurance durations for clients from Kazakhstan:

|  |  |  |
| --- | --- | --- |
| № | [Длительность страхования] Duration of insurance in days | Count |
| 1 | 365 | ? |
| 2 | 183 | ? |
| 3 | 184 | ? |
| 4 | 363 | ? |
| 5 | 343 | ? |

H.6 Analyze the cities and seasons when most accidents occurred

Please, fill in the following table for Almaty and Nur-Sultan:

|  |  |  |  |
| --- | --- | --- | --- |
| № | City | Seasons | Number of accidents |
| 1 | Almaty | Summer | ? |
| Autumn | ? |
| Winter | ? |
| Spring | ? |
| 2 | Nur-Sultan | ? | 3 |
| ? | 4 |
| ? | 6 |
| ? | 2 |

H.7 Have citizens of other countries or right-hand drive cars been in accidents, and if so, what is their ratio compared to all the data?

12) Their relationship is ? (round up 2 digits after floating point)

I) Correlation between variables

I.1 Find the dependence of data on KBM, gender, and city in relation to the price of insurance

13) Hint: You need to use Pearson correlation method. (round up 2 digits after floating point)

Correlation between KBM and gender is ?. KBM and city is ?. KBM and [Страховая Премия] Insurance Premium is ?

I.2 Do we have a strong correlation between Driving experience and age? Check this point statistically (statistical test) (round up 2 digits after floating point)

14) Coefficient of Pearson’ Correlation is ?.

There is a statistically significant correlation between driving experience and age. ? (true or false)

Hint: if p-value < 0.05

J) Probability of an event

J.1 What is the probability of an accident in winter (December - February)

15) Probability of accident in winter is ? (round up 2 digits after floating point)

J.2 What is the probability of an accident in winter and in Almaty

16) Probability is equal to ? (round up 3 digits after floating point

K) Draw a graph

K.1 Histogram of distribution by city

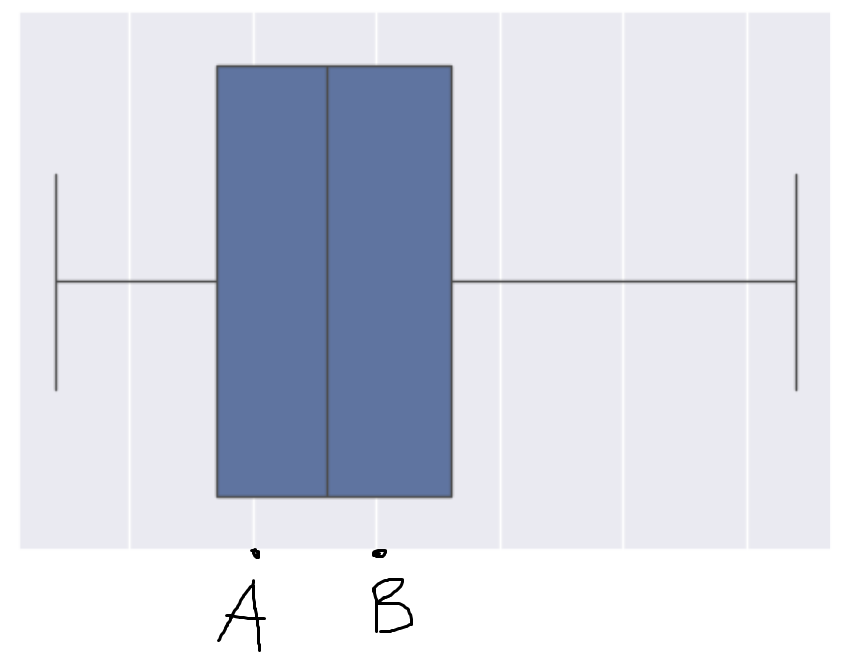
17) Please write Distribution of cars by cities (top 10) in descending order: ?, ?, ?, ?, ?, ?, ?, ?, ?, ?

K.2 Boxplots for age with outlier less than 25% and more than 75%

18) What is the value of A and B in the figure below (approximately)? :

A = ?, B = ?

Hint: you need to write integer number. You are expected to see the below graph after completing step K.2



K.3 Linear graph of variables Driving experience and KBM

19) What is the value of KBM when [Стаж вождения] Driving Experience is equak to 20? Write approximate integer value. ?

What is the value of KBM when [Стаж вождения] Driving Experience is equak to 60. Write approximate integer value. ?

You are expected to see the below graph after completing step K.3.

Изображение выглядит как текст, линия, снимок экрана, График

Автоматически созданное описание

L) Calculate the net profit of the insurance company for 2013, taking into account insurance premiums and losses

20) Net profit of the insurance company for 2013 is ? (please drag-and-drop digit by digit)

That’s all about 1st Assignment. I hope that you really enjoyed during solving these interesting problems and derived some useful information for yourself. Thank you for your accurate reading and for your attention until the end. If you have any questions, please do not hesitate to contact me via MS Teams.

Kind Regards,

Ruslan Omirgaliyev

Senior-Lecturer of Department of Computer Engineering

Astana IT University