**EN:**

**First Data Overview**

A) At the very first stage, load data into the session using Pandas library functions, check the data and rename the columns to more convenient names (without spaces)

**Data Processing**

B) Remove duplicates in the data. Be careful when selecting duplicates - you need to remove complete duplicates.

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

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.

E) Tidy up the values ​​in the [Город] City and [Год выпуска] Year of manufacture columns. For the [Year of manufacture], information about the [Year] is enough, without the day and month.

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.

**Basic Statistics and Data Visualizations**

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

H) Perform basic analytics, answer the following questions:

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

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

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

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

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

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

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?

I) Correlation between variables

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

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)

J) Probability of an event

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

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

K) Draw a graph

K.1 Histogram of distribution by city

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

K.3 Linear graph of variables Driving experience and KBM

**RU:**

**First Data Overview**

A)На самом первом этапе загрузите данные в сессию используя функции библиотеки Pandas, проверьте данные и переименуйте столбцы в более удобные названия (без пробелов)

**Data Processing**

B) Удалите дубликаты в данных. Будьте осторожны при выборе дубликатов – необходимо удалить полные дубликаты.

C) Заполните пустые поля (Missing data) в определенных столбцах значениями (Imputation) – в некоторых случаях необходимо продумать логику.

D) Сконцентрируйте ваш проект только на легковых автомобилях не старше 25-и лет. Исключите автобусы, прицепы и мотоциклы. \* Имейте в виду что датасет 2013 года.

E) Приведите в порядок значения в столбцах Город и Год выпуска. Для года выпуска достаточно информации о Годе, без дня и месяца.

F) Исключите из расчётов записи страховок сроком менее 30-и дней - для них страховая политика формируется другим образом. Убедитесь что при расчёте модели, цена корректная в зависимости от времени страховки.

**Basic Statistics and Data Visualizations**

G) Для каждого из полей, выполните базовую статистику (напр. средние знач., квартили, стандартное отклонение) - сделайте первые выводы на основании результатов

H) Сделайте базовую аналитику, ответьте на следующие вопросы:

H.1 Сколько максимально автомобилей у одного человека

H.2 Найдите соотношение аварий по полу - Мужчины и Женщины

H.3 Какое количество автомобилей купили страховку в наиболее крупных городах Казахстана – Астана (в датасете Нур-Султан), Алматы, Шымкент и Атырау

H.4 Какие 3 модели автомобилей являются наиболее популярными среди Казахстанцев?

H.5 На какой срок в основном покупают страховку Казахстанцы?

H.6 Проанализируйте города и времена года, когда происходило больше всего аварий

H.7 Попадали ли граждане других стран или автомобили с правым рулем в ДТП, и если да, какое их соотношение по сравнению со всеми данными?

I) Корреляция между переменными

I.1 Найдите зависимость данных по КБМ, полу, и городу по отношению к цене страховки

I.2 Есть ли сильная корреляция между стажем вождения и возрастом? Проверьте этот момент статистически (статистический тест) (округлите до 2 цифр после запятой)

J) Вероятность наступления события

J.1 Какова вероятность ДТП в зимнее время (декабрь - февраль)

J.2 Какова вероятность ДТП в зимнее время и в г. Алматы

K) Нарисуйте график

K.1 Гистограмма распределения по городам

K.2 Боксплоты для возраста с нахождением outliear меньше 25% и более 75%

K.3 Линейный график переменных Стаж вождения и КБМ

L) Рассчитайте чистую прибыль страховой компании за 2013 год с учетом страховых премий и убытков