Programming for Economists Practice Exam

You have two hours in total to complete this exam. You should allocate roughly 30 minutes for each applied dataset (the exam contains 4 of these in total).

Numeric input answers should be answered to an accuracy of **one decimal point.** Please do not include trailing zeros; for instance, a value of 3.0 should be reported as 3. Testvision will only recognize answers with the correct granularity, so please mind the decimals.

Part 1: City Budgets (25 points)

Dataset city_budget.xlsx contains data on cities' budgets, service quality, and population. Variables are as follows:

- city_id: Unique identifier for each city
- population: City population
- region: Which region of the country the city is located in
- total_budget: The total fiscal budget for the city, in thousands of euros
- public_safety_budget: The total fiscal budget the city allocates to police and fire, in thousands of euros
- education_budget: The total fiscal budget the city allocates to education, in thousands of euros
- service_quality_index: An index of cities' service quality (0-100)

The following questions concern city_budget.xlsx.

1. Report the mean and median of total_budget, in thousands of euros (2 points).
2. Within each region, compute the mean and median percentage of cities' total budgets spent on education. (8 points)

	North	East	South	West
Mean Education Share	_____%	_____%	_____%	_____%
Median Education Share	_____%	_____%	_____%	_____%

3. Compute the budget that each city spends per person, a 'budget per capita'. Report the top five city IDs by budget per capita. Report the IDs from smallest to largest. (5 points)

4. Create a scatterplot that shows the relation between the service quality index and total city budget. The plot will be graded on the below components. (Total 10 points)
    - The plot is a scatterplot. (1 point)
    - The scatterplot has service_quality_index on the y-axis. (1 points)
    - The scatterplot has total_budget on the x-axis. (1 points)
    - The y-axis should run from 30 up to and including 100. (1 point)
    - The y-axis should be displayed in intervals of 10. (1 point)
    - The x-axis should run from 150000 up to and including 850000. (1 point)
    - The x-axis should be displayed in intervals of 100000. (1 point).
    - The plot contains a line of best fit to the plot. (2 points)
    - The line of best fit has a confidence band. (1 point)

Part 2: Public Transit (25 points)

Dataset public_transport_survey.xlsx contains data on a survey of people's experience with public transportation. Variables are as follows:

- respondent_id: Unique identifier for each city
- age: Respondent's age
- transport_mode: The mode of public transit identified as the primary mode taken by the respondent
- num_trips_week: The number of times per week the respondent reports taking that form of public transit
- satisfaction: The respondent's reported satisfaction with that mode of public transit (1-5 scale)
- punctuality_rating: The respondent's reported rating of the transit mode's punctuality, with higher scores responding to better performance (1-5 scale)

The following questions concern public_transport_survey.xlsx.

1. This dataset contains errors. Identify the percentage of observations in the dataset containing apparent errors in each of the below variables. This is not a trick question; each of the below variables contains errors.
    a. num_trips_week (1 point)
    b. satisfaction (1 point)
    c. punctuality_rating (1 point)

For questions 2-4, replace all apparent errors in the dataset with NAs.

2. Suppose that someone is a 'all-weekday commuter' if they take 10 or more trips with their primary mode of transit each week, and that they are 'high-

satisfaction' if their satisfaction rating is ≥ 4. Report the percentage of high-satisfaction participants for each combination of all-weekday commuter status and primary transit mode, ignoring participants with errors in variable 'satisfaction'. (8 points)

	Bus	Metro	Train	Tram
All-Weekday Commuter	_____%	_____%	_____%	_____%
Not All-Weekday Commuter	_____%	_____%	_____%	_____%

3. Report the mean and median values of punctuality_rating by transit type, ignoring participants with errors in variable 'punctuality_rating'. (8 points)

	Bus	Metro	Train	Tram
Mean Punctuality Rating	_____	_____	_____	_____
Median Punctuality Rating	_____	_____	_____	_____

4. Create a bar plot displaying the average number of trips taken by transport mode. The plot will be graded on the below components.
    a. The plot is a bar plot. (1 point)
    b. The plot indicates the correct transport modes on the x axis. (1 point)
    c. The plot has num_trips_week on the y-axis. (1 point)
    d. The bars should represent the mean number of weekly trips taken by respondents naming the respective transit mode as their primary mode of transportation (1 point).
    e. The y-axis should range from 0 up to and including 12. (1 point)
    f. The y-axis should vary in intervals of 2. (1 point)

Part 3: Hospitals (25 points)

Dataset hospital_data.xlsx contains facility data for a series of hospitals. Variables are as follows:

- hospital_id: A unique identifier for hospitals
- region: The region in which the hospital is located
- bed_count: The number of hospital beds in the hospital
- staff_count: The total number of employees in the hospital
- patient_satisfaction: The average patient satisfaction in the hospital (1-5 scale)
- mortaility_rate: The percentage of admissions who die during their inpatient stay

Dataset readmission_rates.xlsx contains data on readmission rates for the same hospitals. Variables are as follows:

- hospital_id: A unique identifier for hospitals
- thirty_day_readmit_rate: The percentage of all patients who are readmitted to the hospital within 30 days
- heart_failure_readmit: The percentage of all heart failure patients who are readmitted to the hospital within 30 days
- pneumonia_readmit: The percentage of all pneumonia patients who are readmitted to the hospital within 30 days

The following questions concern both hospital_data.xlsx and readmission_rates.xlsx.

1. Report the mean average wait time for each region. (4 points)

	North	East	South	West
Mean Average Wait Time	_____ min	_____ min	_____ min	_____ min

2. Suppose that hospitals are 'high-satisfaction' if the average patient satisfaction for the hospital is ≥ 4. Report the mean wait time and mortality rate by high-satisfaction status. (4 points)

	Mean Average Wait Time	Mean Mortality Rate
High-Satisfaction	_____ min	_____%
Not High-Satisfaction	_____ min	_____%

3. Divide hospitals into terciles by mortality rate. I.e., divide hospitals into the 50 lowest, 50 middle, and 50 highest hospitals by mortality rate. Within each tercile, report the mean total thirty-day readmission rate, heart failure readmission rate, and pneumonia readmission rate. (9 points)

	Mean Total Readmission Rate	Mean Heart Failure Readmission Rate	Mean Pneumonia Readmission Rate
Lowest Tercile	_____%	_____%	_____%
Middle Tercile	_____%	_____%	_____%
Highest Tercile	_____%	_____%	_____%

4. Divide hospitals into terciles by mortality rate. I.e., divide hospitals into the 50 lowest, 50 middle, and 50 highest hospitals by mortality rate. Create a box plot of the staff-to-bed ratio by tercile. The plot will be graded on the following components. (8 points total)

a. The plot is a box and whisker plot. (1 point)
b. The x-axis clearly indicates which box belongs to which tercile. (1 point)
c. The x-axis is titled, indicating what variable is represented in the terciles. (2 points)
d. The ratio of staff members to beds in each hospital are on the y-axis. (2 point)
e. The y-axis spans from 0.65 to 1.25. (1 point)
f. The y-axis varies in intervals of 0.1. (1 point)

Part 4: Housing Survey (25 points)

Dataset housing_survey.xlsx contains data on rental prices, housing information, and incomes. Variables are as follows:

- household_id: A unique identifier for households
- region_type: The kind of region in which the house is located
- household_size: The number of people living in the household

The following questions concern housing_survey.xlsx.

1. Report the percentage of missing values in each of the below variables. (3 points)

	monthly_income	satisfaction_score	utility_costs
Percent Missing	_____%	_____%	_____%

2. Create a histogram of household sizes. Your plot will be graded on the following attributes. (8 points)
    a. The plot is a histogram. (1 point)
    b. The x-axis indicates the values of household sizes. (1 point)
    c. The x-axis varies from 1 to and including 8. (1 point)
    d. The x-axis and the bins vary in intervals of 1. (2 points)
    e. Counts of households are on the y-axis. (1 point)
    f. The y-axis varies from 0 up to and including 70. (1 point)
    g. The y-axis varies in intervals of 1. (1 point)
3. Measure rent burden by computing the percentage of monthly income spent on rent. Report the average value of this percentage for each unique value of household size, ignoring observations with missing values of monthly income. (8 points)

Household Size	Mean Percentage of Income Spent on Rent
1	_____%
2	_____%
3	_____%
4	_____%
5	_____%
6	_____%
7	_____%
8	_____%

4. Within each region type (urban and rural), report the mean values of monthly income, satisfaction scores, and utility, ignoring missing values. (6 points)

	Mean Monthly Income	Mean Satisfaction Score	Mean Utility Costs
Rural			
Urban			