Programming for Economists
Practice Exam 2

Make sure to save your progress regularly using CTRL + S. You have two hours in total to complete this exam. You should allocate roughly 30 minutes per question block on an 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 3.10 should be reported as 3.1, not 3.10. Testvision will only recognize answers with the correct granularity, so please mind the decimals.

Before starting the exam, please run the following four lines of code in order:

- - - - - - - - - - - -

```
setwd(“INSERT YOUR DIRECTORY TO FOLDER ‘Practice Exam 2’ HERE”) #This can also be done manually

library(tidyverse)

source(“tibblehunter.R”)

start_tibblehunter()
```

- - - - - - - - - - - -

You will receive a message in the console that the “Tibble hunter is started!” This is normal. You may now begin the exam.

Part 1: The Library (20 points)

Dataset library_users.xlsx contains data on a city-wide survey concerning library usage. Variables are as follows:

- user_id: Unique identifier for respondents
- neighborhood: Neighborhood type string
- library_visits_month: The number of times the respondent reports visiting a library each month
- hours_read_week: The number of hours per week the respondent reports are spent reading

- uses_wifi: Indicates whether the respondent reports using wifi at the library
- uses_events: Indicates whether the respondent reports attending library events
- satisfaction_rating: Respondent’s satisfaction rating of library services (1-10 scale)

The following questions concern library_users.xlsx.

1. What percentage of the values of the following variables are missing? This is not a trick question; each of the following variables has some missing values. (3 points)

	Library Visits per Month	Hours Spent Reading per Week	Satisfaction Rating
Missing	_____%	_____%	_____%

2. Within each neighborhood type, (a) what percentage of participants use wifi at the library, and (b) what percentage of participants attend library events? (6 points)

	Downtown	Rural	Suburban
Use Wifi @ Library	_____%	_____%	_____%
Attend Library Events	_____%	_____%	_____%

3. One way people deal with missing values is to assume that individuals who did not fill out a question take on a value of zero. (a) Report the mean hours spent reading per week by neighborhood type after removing observations with missing values of hours spent reading per week. (b) Report the mean hours spent reading per week by neighborhood type after imputing all NA values of hours spent reading per week to zero. (6 points)

	Downtown	Rural	Suburban
Removing NAs	_____ hours/week	_____ hours/week	_____ hours/week
Imputing NAs to 0	_____ hours/week	_____ hours/week	_____ hours/week

4. Create a box plot of library visits per month, removing observations with NA values in library visits per month. The plot will be graded on the following attributes (6 points total):
    a. The plot is a box and whisker plot. (1 point)
    b. The x-axis clearly indicates which box belongs to which neighborhood type. (1 point)
    c. The mean number of library visits per month is accurately displayed on the y-axis. (2 points)
    d. The y-axis ranges from 1 to 7 in intervals of 1. (2 points)

Part 2: Parks & Rec (30 points)

Dataset park_features.xlsx contains data on the characteristics of public parks in a city. Variables are as follows:

- park_id: A unique park identifier
- district: City district
- area_sq_m: Total park area in square meters
- num_trees: Estimated number of trees in the park
- has_playground: Indicates whether the park has a playground
- avg_daily_visits: The average number of visitors to the park each day

Dataset maintenance_reports.xlsx contains data on maintenance reports for the parks in park_features.xlsx. Variables are as follows:

- park_id: A unique park identifier
- cleanliness_score: Inspector rating of park cleanliness (1-10 scale)
- safety_score: Inspector rating of safety (1-10 scale)
- last_inspection_days: Number of days since last inspection

The following questions concern park_features.xlsx and maintenance_reports.xlsx.

1. Report the mean and minimum cleanliness score by district. (8 points)

	Greendale	Hilltop	Lakeside	Riverview
Mean	_____	_____	_____	_____
Minimum	_____	_____	_____	_____

2. Compute the tree density as the number of trees in each park per square kilometer. Report the maximum number of trees per square kilometer for each district. (4 points)

	Greendale	Hilltop	Lakeside	Riverview
Maximum Trees Per Square Kilometer	_____ trees/square km	_____ trees/square km	_____ trees/square km	_____ trees/square km

3. Suppose that a park is 'high-visitation' if its visitorship is above the median across all parks in the city. Within each district, (a) compute the percentage of high-visitation parks amongst parks with playgrounds, and (b) compute the percentage of high-visitation parks amongst parks without playgrounds. (8 points)

	Greendale	Hilltop	Lakeside	Riverview
With Playgrounds	_____%	_____%	_____%	_____%
Without Playgrounds	_____%	_____%	_____%	_____%

4. Create a scatter plot showing the relationship between the amount of time since the last inspection and safety scores. (9 points)
    a. The plot is a scatter plot. (2 points)
    b. The plot has days since last inspection on the x-axis. (1 point)
    c. The x-axis ranges from 0 to 60 in intervals of 15. (2 points)
    d. The plot has safety scores on the y-axis. (1 point)
    e. The y-axis ranges from 4 to 9 in intervals of 1. (2 points)
    f. The plot has a line of best fit with no confidence band. (2 points)

Part 3: Job Training (20 points)

Dataset training_participation_survey.xlsx concerns a survey fielded amongst 350 adults who recently completed or dropped out of publicly funded job training programs. Variables are as follows:

- participant_id: Unique ID for respondents
- program_status: Should take values "Completed and Passed", "Completed but Failed", or "Dropped Out"
- age: Age of respondent

- training_hours: Total hours of training received
- job_offer_within_3mo: Indicates whether the respondent received a job offer within three months
- satisfaction_score: Participants' satisfaction with the program (1-5 scale)

The following questions concern training_participation_survey.xlsx.

1. This dataset contains errors. What percentage of values for each of the following variables contain apparent errors? This is not a trick question; every one of the below variables contains some errors. (3 points)

	Program Status	Satisfaction Score	Training Hours
Error Rate	_____%	_____%	_____%

For questions 2-4, correct clear spelling errors in program status to map them to appropriate values of that variable. Impute apparent errors in satisfaction scores and training hours with the median of those respective variables amongst participants with non-error values of those respective variables.

2. Within each program status, (a) compute the mean satisfaction score for participants who received a job offer within three months, and (b) compute the mean satisfaction score for participants who did not receive a job offer within three months. (6 points)

	Completed and Passed	Completed but Failed	Dropped Out
Received Job Offer Within 3 Months	_____	_____	_____
Received No Job Offer Within 3 Months	_____	_____	_____

3. Bin ages into categories of 18-25, 26-45, and 46-65. Within each age bin, compute the proportion of participants who completed and passed. (3 points)

	Ages 18-25	Ages 26-45	Ages 46-65
Complete + Pass Rate	_____%	_____%	_____%

4. Create a histogram of training hours. Your plot will be graded on the following components (8 points total):
    a. The plot is a histogram. (2 points)
    b. Training hours are displayed on the x-axis in bins of width 30. (2 points)
    c. The count of participants falling into each bin is accurately displayed on the y-axis. (2 points)
    d. The y-axis scales from 0 to 120 in intervals of 20. (2 points)

Part 4: Stocks (20 points)

Dataset mamaa.xlsx contains daily stock data on the five MAMAA tech stocks, specifically Meta (META), Amazon (AMZN), Microsoft (MSFT), Apple (AAPL), and Alphabet (GOOG). The data is organized at the stock-day level, where a single row represents a single trading day for a particular stock. Variables are as follows:

- open_price: The price at which the stock first traded at the beginning of the trading day
- high_price: The highest price at which the stock traded over the trading day
- low_price: The lowest price at which the stock traded over the trading day
- closing_price: The price at which the stock last traded at the end of the trading day
- trading_volume: The number of shares of the stock that were exchanged over the trading day
- ticker: The stock's symbol
- date: The trading day

The following questions concern mamaa.xlsx. As a reminder, the formula for a percentage return is (endprice/startprice - 1) x 100.

1. Across all stock-days in the data, compute the minimum, 25th percentile, median, 75th percentile, and maximum daily trading volume for MAMAA stocks in 2024. (5 points)

Minimum	25th Percentile	Median	75th Percentile	Maximum
_____	_____	_____	_____	_____

2. Based on the previous answer, subset all stock-days in 2024 by whether they are (1) below the 25th percentile, (2) greater than or equal to the 25th percentile or less than or equal to the 75th percentile, or (3) greater than the 75th percentile in

trading volume. Compute the within-day return as the percentage point change between the stock's opening price and closing price on that day. (E.g., if the stock's price is 30% higher at the end of the day, compute it as 30. If the stock's ending price is only 60% of the value at the start of the day, compute it as -40.) Report the maximum and minimum daily return within each of the three trading volume subsets. (6 points)

	Below 25th Percentile in Volume	Between 25th-75th Percentile in Volume	Above 75th Percentile in Volume
Minimum Return	_____%	_____%	_____%
Maximum Return	_____%	_____%	_____%

3. Suppose one purchased each stock at the open price on the earliest available day in 2024 and sold each stock at the closing price on the latest available day in 2024. Compute the return in percentage points for each stock and report their annual return. E.g., if the stock's price is 30% higher at the end of 2024, report 30. If the stock's ending price is only 60% of the value at the start of the year, report -40. (5 points)

	META	AMZN	MSFT	AAPL	GOOG
Return	_____%	_____%	_____%	_____%	_____%

4. Create a line graph showing the progression of closing prices for Microsoft stock for trading days in 2024. Your plot will be graded on the following components. (4 points total)
   a. The plot is a line graph. (1 point)
   b. The closing price of MSFT stock is on the y-axis. (1 point)
   c. The y-axis scales from 350 to 475 in intervals of 25. (1 point)
   d. The date is on the x-axis. (1 point)