DATA SCIENCE 141

A1: Take home assessment 2025

IMPORTANT INFORMATION

Timing and due date

The assessment data and questions are available from **Wednesday 27 August 2025 at 07:00**. You should work on this assessment in your own time. Completing the assessment should not take much more than three to four hours of your time **if** your R skills are good, but please allow yourself much more time than this and do not leave it until the last minute. Make provision for the fact that the assessment could potentially take you much longer than three hours to complete if you struggle with R and must consult notes/help files or obtain help from a teaching assistant/lecturer. You should also allow for the fact that you can run into unforeseen computer problems or encounter errors and in this case will need sufficient time to deal with it. Unexpected events such as illness or other commitments might also impact the time you have available to work on the assessment. You should therefore not leave the assessment until the last day. **The due date for the assessment is Monday 22 September 2025 at 12:00 (midday).** Please note that very strict late submission penalties will be applied (refer to the module framework for details of these) and **no** exceptions will be made. Also remember that submission of A1 is compulsory, and that you therefore cannot pass the module if you do not submit this assessment.

Required submissions

For the assessment, you will have to work in R Markdown. You must submit your Markdown file (.Rmd) as well as the rendered Word version (.docx) thereof; both these files must be uploaded to EMSLearn. Note that for the EMSLearn submissions, only a single attempt per document is allowed. This means that, once you have submitted your assessment files via EMSLearn, you will NOT be able to go back and change the uploaded documents. It is therefore your responsibility to make sure that you are satisfied that everything is complete and correct before you submit your files on EMSLearn. Please also note that after **uploading** the files, you are required to **submit** the files too (you will see a Submit button after the files have been uploaded); if not, the files will be regarded as Draft Submissions on EMSLearn and the assessment will not be marked as complete. **Please double check the status of your submission before the deadline to avoid late submission penalties!**

Help available

Detailed instructions are provided in this document for the completion of important preliminary steps (such as importing the data and taking a sample of the data). Help can be sought from the lecturer or teaching assistants with the *preliminary steps* if required – but within reason. If any of the *assessment questions* are unclear, clarification can be sought from the lecturer on the module discussion forum on EMSLearn only. Specific help (such as assistance with the correct R code to use) will not be provided. All of the necessary skills to enable you to successfully complete the assessment were covered in the practical assignments and tutorials.

Additional information

All data analysis and visualisation must be done in R. Plots and summaries created using other software packages will not be marked.

PLAGIARISM

This take-home assessment constitutes one of the formal assessments for the module Data Science 141. While you are allowed to work on the assessment in your own time and may refer to notes and other sources as described below, it is important that you must work as an individual and may not obtain help from any individuals or sources other than those specified below. **You may also not upload the assessment paper/questions or data online and may not ask for help with the assessment on any online forums either – this implies that you also may not use ChatGPT or other AI sources to assist you with the answering of the assessment questions.**

Your assessment submissions should be entirely your own work. You will be required to submit a signed plagiarism declaration (see Appendix 1 to this document; also available as a stand-alone form on EMSLearn). The signed plagiarism declaration can be uploaded to EMSLearn or a printed copy can be submitted to the lecturer in person by the due date. **Failure to submit a signed plagiarism declaration form means that your assessment will not be marked.**

Please take note that contravention of the academic honesty policy could lead to disciplinary action.

Allowed sources:

- Class notes.
- Textbooks and similar material.
- Online sources such as R help files.
- Teaching assistants/lecturer (help with preliminary steps only).
- Module discussion forum (for clarification of assessment questions only).

What is not allowed?

- You are not allowed to seek help from anyone with any of the assessment questions, including your classmates, teaching assistants or lecturer.
- You are not allowed to seek direct help online. This means that you may not post the assessment data set or any of the questions online on public forums in order to obtain assistance.
- You may not use ChatGPT or similar AI sources to directly generate code and/or answers to the assessment questions.

ADVICE/TIPS FOR SUCCESSFUL COMPLETION OF THIS ASSESSMENT

1. Start by carefully reading through this whole document – including the questions – before you start working with the data.

2. Once you have downloaded the data set and imported it into RStudio, you should spend sufficient time exploring the data before you start working on the assessment.

3. The preliminary steps are important, as they are there to ensure that you are doing the basics right before you start answering the questions. If you struggle with any of these preliminary steps, please ask for help as soon as possible. Remember that it is recess from 6 – 14 September 2025, so there will be no help available during that period. The discussion class on 26 August 2025 was used to explain the basics of the assessment, and all subsequent practical Q&A sessions and tutorial sessions can be used as opportunities to ask for help with the preliminary steps in person.

4. In the discussion type questions, make sure that you think about the context of the data. Remember that data science is more than just producing plots and building models. If you create plots, think carefully about the information you want to convey in your plot and how best to achieve that. You are not being assessed purely on your practical R skills, but also on how you can "think" with data.

5. Many of the assessment questions do not necessarily have only one correct answer. There can be several suitable ways to visualise certain data for instance; in terms of the discussion, your own interpretation/opinion is important. Therefore, make sure that you are clear in describing any conclusions you draw from the data and your analyses thereof, and always substantiate your answers with numbers and/or plots as required.

6. Use suitable headings so that it is clear what question is answered by each piece of code and output. Interpretations of plots and statistics should follow after the relevant code and output. If it is not clear what question is being answered by the code/output, it will not be marked. **Your rendered document should include all relevant code, output and interpretation/discussion** (unless explicitly directed otherwise).

7. Think carefully about the output you want to produce with your code. It should be clear from the rendered version of the file what the answer to a particular question is. You shouldn't produce lots of output and expect the grader to find the relevant piece of output to answer a question. As an example, if you are asked to calculate an average value, it will not be sufficient to simply produce the output of the `summary()` function to answer the question. Be purposeful about the code you are writing, and make sure you give sufficient interpretations and discussions.

8. Make sure your plots have suitable titles and labels (including for the axes where relevant). You should also pay attention to the interpretability and visual appeal of your plots. Awarding of marks for plots will take this into account.

START OF ASSESSMENT

Background scenario

New York City (NYC) in the United States of America is served by three main airports: John F Kennedy International Airport (JFK), LaGuardia Airport (LGA) and Newark Liberty International Airport (EWR). While LaGuardia Airport mostly handles domestic flights (that is, flights travelling within the borders of the USA), JFK and Newark airports handle both domestic and international flights. All three of these airports are very busy; in fact, the New York metropolitan area has the busiest airport system in the USA and the second busiest in the world (after London in the United Kingdom).[1]

Of the three airports, JFK is the largest and also the busiest: in 2024, it served 63.3 million passengers, of which 44% travelled on domestic flights. Newark served 48.9 million passengers (69% travelling on domestic flights), and LaGuardia 33.5 million (95% on domestic flights).[2]

An official report[3] puts the number of scheduled domestic flights for 2023 for these three airports at 565 263. The R package `nycflights23` contains data for flights departing from these three airports in 2023. According to the package documentation[4], it contains information about all flights that departed from these three airports in 2023. However, an investigation of the provided data suggests that only domestic flights are included in the package data. Furthermore, the package contains data for only 435 352 flights, which is substantially fewer than the number contained in the 2023 *Airport Traffic Report of the Port Authority of New York and New Jersey*[3]. Nevertheless, it is still a vast data resource with a wealth of information on flight records.

[1] https://en.wikipedia.org/wiki/Aviation_in_the_New_York_metropolitan_area
[2] https://www.panynj.gov/content/dam/airports/statistics/statistics-general-info/annual-atr/ATR_2024.pdf
[3] https://www.panynj.gov/content/dam/airports/statistics/statistics-general-info/annual-atr/ATR_2023.pdf
[4] https://cran.r-project.org/web/packages/nycflights23/index.html

Within the `nycflights23` package, there are five data sets:

1. The main data set `flights`, which contains flight records such as origin and destination airport codes, scheduled and actual departure and arrival times and dates, carrier codes, airplane identifiers (known as tail numbers), time spent in the air as well as distance travelled.
2. An `airlines` data set, containing airline names, which can be matched to the carrier codes in the `flights` data set.
3. `airports`, which contain airport names (corresponding to the airport codes used in the origin and destination codes in `flights`).
4. The `planes` data set, which contains an airplane identifier (tail number) as well as information such as manufacturer and model for each plane.
5. A `weather` data set containing hourly meteorological data for all three NYC airport locations for the entire year.

These five data sets were used to construct a single data set, including some derived variables. A couple of observations from the original `flights` data set were also deleted. The resulting data set contains 435 350 observations and 31 variables.

You – as a data scientist – have been tasked to help answer certain questions about domestic flight departures in New York City in 2023, using the provided data set.

Note that you are welcome to examine the original data sets in the `nycflights23` package should you wish to. However, it is not necessary to do so in order to complete this assessment (unless you are specifically instructed to do it for a specific question). All assessment questions should be answered using the provided data set (A1_data.csv) on EMSLearn.

The data

The variables contained in the assessment data set in the file A1_data.csv are as follows:

Variable name	Variable description
`row_id`	Unique row identifier
`year`	The year of the flight. For this data set therefore, you should have `year = 2023` for all observations.
`month`	The month in which the flight was scheduled to depart.
`day`	The day on which the flight was scheduled to depart.
`date`	This is a variable derived from the `year`, `month` and `day` variables, to give a date representation in a single variable.
`dep_time`	Actual departure time of a flight. Missing (NA) values for this variable indicate flights that were cancelled.
`sched_dep_time`	This is the time at which a flight was scheduled to depart.
`dep_delay`	Difference between actual and scheduled departure time of a flight.

`Departure`	This variable takes on three possible values:
	`On Time`: Flights with `dep_delay = 0`
	`Early`: Flights with `dep_delay < 0`
	`Late`: Flights with `dep_delay > 0`

Variable name	Variable description
arr_time	Actual arrival time of a flight. Cancelled flights or flights that had to return to the origin airport will not have a value for this variable.
sched_arr_time	Scheduled arrival time of a flight.
arr_delay	Difference between actual and scheduled arrival time of a flight.

Arrival	This variable takes on three possible values:
	`On Time`: Flights with `arr_delay = 0`
	`Early`: Flights with `arr_delay < 0`
	`Late`: Flights with `arr_delay > 0`

carrier	This is a code assigned by the International Air Transport Association (IATA) and is used to identify an airline (carrier).
flight	A numeric code to identify a flight. This is usually used in conjunction with the carrier code to identify a route. Note that this number is not unique; for instance, the flight number 69 in conjunction with the carrier code **DL** refers to the Delta Airlines flight from JFK to Honolulu. However, the flight number 69 in conjunction with the carrier code **UA** refers to the United Airlines flight from Newark to Palm Beach.
tailnum	This is a way of uniquely identifying a specific aircraft (almost akin to the VIN of a motor vehicle). This field therefore indicates which particular aircraft was used for the given flight.
Plane_manuf	The manufacturer of the aircraft used for the particular flight. This variable can take on four possible values: Airbus, Boeing, Bombardier or Embraer. Missing values for this field can either indicate that a flight was cancelled, or that aircraft information for the particular flight is not available.
Plane_model	The model of the aircraft used for the flight. Missing values can either indicate a cancelled flight, or missing aircraft information.
origin	The airport from which the flight departed/was scheduled to depart. For this data set, it can only have 3 possible values: **JFK** (John F Kennedy International Airport), **EWR** (Newark Liberty International Airport) or **LGA** (LaGuardia Airport).
dest	The destination airport for the flight. This is a three-character alpha-numeric code issued by the U.S. Department of Transportation, which is used as official designation for the airport
air_time	The airborne hours of an aircraft for a particular flight, computed from the moment the aircraft leaves the ground until it touches the ground at the end of the flight. This will therefore not be equal to the difference between `arr_time` and `dep_time`, as these two variables consider the time an aircraft left the gate and returned to the gate, and not actual lift-off and touchdown times.
distance	Distance (in miles) between departure and destination airports.
time_hour	A variable giving the date of the flight as well as the hour of the scheduled departure time. Together with the origin airport of the flight, this can be used to look up weather data for each flight.
temp	Temperature, in Fahrenheit.
dewp	Dewpoint temperature, in Fahrenheit.
humid	Relative humidity.
wind_dir	Wind direction (in degrees).
wind_speed	Wind speed (in miles per hours)
wind_gust	Wind gust speed (in miles per hours)
precip	Precipitation (in inches)
visib	Visibility (in miles)

Preliminary steps

These steps are important and should be followed carefully. Pay special attention to what the different files and objects should be named. If you don't follow the prescribed naming conventions, it could result in your assessment not being marked, or forfeiting of marks for some questions.

If you struggle with any of these preliminary steps, you are allowed to seek assistance from the lecturer (during practical Q&A sessions) or from a teaching assistant (during practical or tutorial sessions).

a) The file `A1_data.csv` contains the data required for this assessment. The file is available on EMSLearn under the **Assessment 1** topic. Download this file and save it to the computer you will be working on. Make a note of the location where you saved the file, as you will need to access it later.

b) Open RStudio and create a new R Markdown file. This file should be a **Document**, and the title of the file should be **A1 submission**. The **Author** name should be your full name and student number. In the date field, type only the year (2025). Make sure that you choose **Word** as your output format.

c) In the markdown file you created in Step (b), delete everything from `## R Markdown` onwards. In other words, leave your YAML header in place as well as the first code chunk named `setup`; everything else can be deleted.

d) In your markdown file, create a new code chunk. In this code chunk, write code to load the `tidyverse` group of packages, as well as the `tree` package. **Include code that suppresses any output or messages from this specific code chunk in your final document.**

e) Before you continue, **Knit** your markdown document. Save it in the same location where you saved the `A1_data.csv` file. When prompted to provide a file name, the format of the name you provide should be `Surname_studentnumber`. For instance, if your surname is Smith and your student number is 12345678, you should name your file `Smith_12345678`.

f) Create another new code chunk below the one you created in Step (d). Within this code chunk, write code to import the data set you downloaded in Step (a) into RStudio and store it in an object called **`main_data`**. Note that the data set contains the variable (column) names in the first row, so you should include the argument **`header = TRUE`** in your **`read.csv()`** function.

g) Check the dimensions of the data set you imported in Step (f) – you should have 435 350 observations and 31 variables.

h) Check the variable types (remember that you can use the **`str()`** function to do this). After importing, the variable types should correspond to the types indicated in the table on the next page. If any of the variable types are incorrect, you will have to write code to fix this.

Variable name	Variable type
row_id	int
year	int
month	int
day	int
date	Date
dep_time	int
sched_dep_time	int
dep_delay	int
Departure	Factor
arr_time	int
sched_arr_time	int
arr_delay	int
Arrival	Factor
carrier	Factor
flight	int
tailnum	chr
Plane_manuf	Factor
Plane_model	Factor
origin	Factor
dest	Factor
air_time	int
distance	int
time_hour	Date
temp	num
dewp	num
humid	num
wind_dir	int
wind_speed	num
wind_gust	num
precip	num
visib	num

i) You will need to take a sample (without replacement) of the observations in the data set before you start working on the assessment questions. To ensure reproducibility of your work, you need to set a seed before writing the code to draw a sample. ***The seed should be set as the last digit of your student number.*** For example, if your student number is 12345678, you should use the command `set.seed(8)`.

j) Draw a sample of size 400 000 from the original data set. Store this sample in an object called **`test_data`**. Make sure that you set **`replace = FALSE`** for a sample without replacement. (You can refer to the instructions from Practical Assignment 4 if you are unsure of how to draw a sample.)

k) To check whether your sample is correct, please sort the records in the **`test_data`** object ascending according to `row_id` and compare the 1st, 100th, 200th and 500th `row_id` values in your sorted **`test_data`** object to the `row_id` values in the relevant row on the Seed check.pdf document provided on EMSLearn. If your

sample appears to be incorrect, please check where you went wrong and correct it, or reach out to the lecturer or a teaching assistant for help. Once you are satisfied that these steps have been followed exactly and that you are working with the correct data set, you can proceed with answering the assessment questions.

The allocated marks for the preliminary steps will take the following into account:

- Correct output format
- Correct header
- Packages correctly loaded
- Data correctly imported
- Correct variable types
- Sample correctly drawn
- Markdown file runs correctly (i.e. reproducible)

Marks for preliminary steps: 10

****

In addition to the marks for the preliminary steps, you will also be awarded a mark out of 10 for "overall impression". This mark will take into account the overall neatness of your rendered document, whether unnecessary output was excluded, the aesthetical aspects of your data visualisations as well as spelling/grammar/clarity of interpretations and explanations.

Overall impression: 10

****

Note that, since the provided data set was derived and modified from the original (and you will also be working with a sample of this data), you will not be able to compare your answers to any analysis of the original data which you might be able to find online!

Questions

This constitutes the formal assessment portion. Please remember that you are not allowed to seek help from anyone with any of these questions, including your classmates, teaching assistants or lecturer. Your assessment submission should be entirely your own work.

Please be purposeful about the code and output you include in your rendered Word document. Marks will be deducted for unnecessary output. For example, printing the entire data set in your rendered document will be considered unnecessary output.

For questions 1 – 8 below, use the sample you created in preliminary step (j); in other words, work with the `test_data` object.

1. This question considers the `distance` variable.

1.1 Calculate the mean of `distance`.

(1)

1.2 Calculate the median of `distance`.

(1)

1.3 Create a suitable plot to show the distribution of the `distance` variable. Based on this plot, how would you describe the distribution?

(5)

1.4 Would you say there are any outliers in the `distance` variable? Substantiate your answer.

(3)

2. Create a bar chart showing the number of scheduled flights in the data set for each category of the `Plane_manuf` variable. Display the number of flights in each `Plane_manuf` category on the individual bars of your bar chart as well. (Note that none of your practical assignments or tutorials specifically considered adding labels such as these to a bar chart. You will therefore have to do research online to figure out how to do it!)

(5)

3. Which New York airport was the busiest in 2023 (in terms of the number of scheduled departures)? Create a suitable table to substantiate your answer.

(4)

4. Not all destinations are directly connected to all three New York airports. For instance, there are no (direct) flights from LaGuardia to Albuquerque International Sunport (airport code ABQ). This airport is therefore only connected to JFK and Newark Airport.

4.1 Which of the three New York airports connects to the largest unique number of destinations? Note that you should not consider the number of flights to each destination, but instead you should only consider the number of destinations that are connected to each origin airport.

(2)

4.2 Give the top 3 destinations (in terms of the number of scheduled departures to each destination) for each of the three New York airports.
(Note that the provided data set only provides FAA codes for each destination. However, you can look at the `airports` data set in the `nycflights23` package, which contains the airport names corresponding to each FAA code.)

(6)

5. Identify the quietest day of the year in New York in terms of the number of scheduled departures (that is, the day with the lowest number of scheduled departures). Why do you think this particular day was the quietest day?

(3)

6. The provided data set contains carrier codes, which correspond to different airlines. Which three airlines had the largest number of scheduled flights in 2023 (across all New York airports combined)? You should give the name of the airline, not just the carrier code.
(Hint: The `airlines` data set in the `nycflights23` package contains the carrier codes as well as the corresponding airline names. You can therefore use this data set to look up the names of the carriers you identify in this question.)

(3)

7. Scheduled departure time (`sched_dep_time`) is represented as a continuous number in the data set, although it actually represents a time. For instance, 1338 is used in the data set to represent the time 13H38, or 1:38 pm.

The %/% operator in R performs integer division and therefore gives the *floor* of a number; that is, while `13 / 5 = 2.6`, integer division gives `13 %/% 5 = 2`. The %% operator gives the remainder. Therefore, `13 %% 5 = 3`. (You can look at the help files for these operators if you would like more information.)

7.1 Use these modular arithmetic operators to transform the scheduled departure time variable into separate hour and minute variables. You should name these two new variables `dep_hour` and `dep_min`. For instance, the time 13H38 (which is represented in the data set by 1338) should have a value of 13 for the `dep_hour` variable and a value of 38 for the `dep_min` variable.

(2)

7.2 Use a line graph to show the number of flights (for all three airports combined) that were scheduled to depart at each minute past the hour (over the entire year). Your line graph should therefore have `dep_min` on the x-axis and the number of flights on the y-axis. What does this plot suggest about scheduled departure times?

(6)

8. The `air_time` variable contains missing values. These values represent flights that were cancelled and therefore never actually departed. These missing values may also represent some other anomalies in the data.

8.1 Calculate the number of flights with missing `air_time` values in the data set.

(1)

8.2 Create a new data set named `departed`. This data set should exclude all flights with missing values for the `air_time` variable. It should contain all the original variables in the `test_data` object (including any new variables created in the previous questions).

(2)

For questions 9 – 13 below, use the data set you created in Question 8.2; in other words, work with the `departed` data set.

9. Create a suitable plot to illustrate the relationship between departure airport (`origin`) and departure status (`Departure`). Based on this plot, which airport would you say has a better track record in terms of timely departures?

(5)

10. Create side-by-side boxplots to show how the distribution of flight duration (`air_time`) differs according to the departure airport. Based on the plot, comment on the difference in flight duration among the three New York airports.

(4)

11. Calculate the average flight duration (`air_time`) per departure airport.

(2)

12. In this question, the relationship between departure delays and arrival delays will be considered.

12.1 Calculate the correlation between `dep_delay` and `arr_delay`. Round your answer to three decimal places.

(2)

12.2 Create a scatterplot showing `dep_delay` on the x-axis and `arr_delay` on the y-axis.

(3)

12.3 Based on your answers in 12.1 and 12.2, comment on the strength and nature of the relationship between `dep_delay` and `arr_delay`. Also comment on the possible reason(s) for this relationship.

(4)

13. The `Departure` and `Arrival` variables were created based on the `dep_delay` and `arr_delay` variables. Any flight that arrived or departed even 1 minute late was coded as a `Late` flight. Airport management feels that this does not give an accurate reflection of the true state of affairs, and that very minor delays are acceptable (and even the norm). They therefore ask you to examine the data and answer the following questions:

13.1 How many flight departures are delayed by less than 10 minutes?

(1)

13.2 How many flight arrivals are delayed by more than 15 minutes?

(1)

13.3 How many flights departed more than 10 minutes late but made up time and arrived earlier than the scheduled arrival time?

(2)

13.4 Create a new categorical variable called `major_delay`. This variable should have the value `Yes` if a flight arrived more than 15 minutes after its scheduled arrival time, and a value of `No` otherwise.

(4)

13.5 Use the `tree()` function in R to fit a tree to predict whether a flight experienced a major delay (in other words, the target variable should be your newly created variable `major_delay`). Only the weather-related variables (`temp`, `dewp`, `humid`, `wind_dir`, `wind_speed`, `wind_gust`, `precip` and `visib`) should be used as potential attributes for the tree. Plot the tree and also print summary output for the tree.

(5)

13.6 Based on your output in Question 13.5, would you say weather had any influence on flight delays? Substantiate your answer.

(3)

Marks for questions: 80

TOTAL MARKS: 100

Before you submit:

- Make sure that your rendered Word document contains headings to indicate the different question numbers.
- Check that your rendered Word document does not include unnecessary output. For instance, including a print-out of the entire dataset will be considered unnecessary output! **Marks will be deducted for this.**

Remember that you should submit the following documents:

1. RMarkdown file, in the format Surname_studentnumber.Rmd. (Uploaded to EMSLearn.)
2. Rendered Word version of the RMarkdown file, in the format Surname_studentnumber.docx. (Uploaded to EMSLearn.)
3. Signed plagiarism declaration. (Scanned copy uploaded to EMSLearn or physical copy handed in to the lecturer by the due date.)

Remember that you have to Submit your files on EMSLearn after uploading them!

APPENDIX 1

Plagiarism declaration

1. Plagiarism is the use of ideas, material and other intellectual property of another's work and to present it as my own.

2. I understand what plagiarism is and I am aware of Stellenbosch University's policy in this regard.

3. I agree that plagiarism is a punishable offence because it constitutes theft.

4. I also understand that direct translations are plagiarism.

5. All quotations and contributions from any sources whatsoever (including the internet) have been cited fully. I understand that the reproduction of text without quotation marks (even when the source is cited) is plagiarism.

6. I declare that the work contained in this assessment is my own original work and I have not obtained help from anyone. I acknowledge that copying someone else's assessment and / or code, or part thereof, is wrong, and that submitting identical work to others constitutes a form of plagiarism.

7. I declare that I have not permitted anyone else to copy my answers and / or code or provided help to another student in completing this assessment.

8. I have not copied this assessment or created images of the screen showing any part of the assessment (including, but not limited to, images obtained by means of screenshots, photographs and any software applications). I have also not shared the data or any part of the assessment questions with anyone else, including uploading it online.

First name: …….

Surname: ……

Student number: ……

Date: ……

Signature: ……