Assignment 1 (15 marks)

The assignment is based on two datasets:

International Airlines - Traffic by city pairs (city_pairs.csv)

This dataset contains information about international flights to and from Australia in monthly intervals from January 1985 until September 2022. The data is provided by the Australian Government and covers passenger, freight, and mail carried between city pairs connected by single flight number services.

https://data.gov.au/data/dataset/international-airlines-traffic-by-city-pairs/resource/ebcafd83-9514-4f72-a995-fe7ee90cb9da

International Airlines - Operated Flights and Seats (seats.csv)

This dataset contains information about international airlines to and from Australia in monthly intervals from September 2003 until September 2022. The data is provided by the Australian Government and covers airline activity, airport locations, and maximum seat capacity.

https://data.datahub.freightaustralia.gov.au/dataset/international-airlines-operated-flights-and-seats

-

You are expected to manually inspect the datasets and answer the questions. Attributes are self-explanatory or elaborated below further. Please ask for clarification on Webcms if needed.

The data has been pulled directly from public sources, and we are not responsible for political correctness or other data inaccuracies as this is purely a learning exercise for data services engineering.

The output and marking scheme are included with each question, and penalties are given for issues such as incorrect values, row/column names, logic, errors, etc. Marks are awarded independently for each question – errors do not carry through between questions.

Question 1 (1.5 Marks)

Description: Using the "city_pairs.csv" dataset, load it in a data frame and add three new columns to it. The values should be one of "IN", "OUT", or "SAME", depending on whether inbound or outbound numbers are higher, or they are the same:

"passenger_in_out" – whether "Passengers_In" or "Passengers_Out" passenger numbers are higher, or otherwise the same

"freight_in_out" - whether "Freight_In_(Tonnes)" or "Freight_Out_(Tonnes)" freight numbers are higher, or otherwise the same

"mail_in_out" - whether "Mail_In_(Tonnes)" or "Mail_Out_(Tonnes)" mail numbers are higher, or otherwise the same

Output: a data frame with all the columns you have loaded from city_pairs.csv with three new columns. These should ~~both~~ have the added three columns "passenger_in_out", "freight_in_out", and "mail_in_out".

Marks:

0.5 mark – loaded data frame correctly

1 mark – all three column values correct

Penalties of -0.25 mark per error are given.

Question 2 (1.5 Marks)

Description: Using the data frame from Question 1, create a new data frame with unique “AustralianPort” values and their total count of the “IN” or “OUT” labels derived in Question 1 for passengers, freight, and mail since January 1985 in separate columns. This must be sorted in descending order by "PassengerInCount” (highest to lowest). There is no need to include counts for “SAME”.

“PassengerInCount” - count of “IN” for passengers for a unique “AustralianPort”

“PassengerOutCount” - count of “OUT” for passengers for a unique “AustralianPort”

“FreightInCount” - count of “IN” for freight for a unique “AustralianPort”

“FreightOutCount” - count of “OUT” for freight for a unique “AustralianPort”

“MailInCount” - count of “IN” for mail for a unique “AustralianPort”

“MailOutCount” - count of “OUT” for mail for a unique “AustralianPort”

Output: a data frame (7 columns) with a row for each unique “AustralianPort” value, and the following columns: “PassengerInCount”, “PassengerOutCount”, “FreightInCount”, “FreightOutCount”, “MailInCount”, “MailOutCount”.

Marks:

0.25 mark – each column with fully correct values and column name

The 0.25 mark is not awarded for a column if the values or column name is wrong

Question 3 (2 marks)

Description: Using the data frame from Question 1, create a new data frame with unique “Country” values and their average monthly passenger, freight, and mail values (including 2dp) for their respective "_In” and “_Out” columns separately. This must be sorted in ascending order by “Passengers_in_average” (lowest to highest). If a country does not appear in a month, they need to be considered as 0 and still contribute to the average.

Take the following example input data frame which is from Jan 2022 to Feb 2022:

Month	AustralianPort	ForeignPort	PassengerIn	PassengerOut	Country	...
Jan 2022	Sydney	Shenzhen	10	20	China	
Jan 2022	Brisbane	Toronto	5	10	Canada	
Jan 2022	Melbourne	Singapore	8	12	Singapore	
Feb 2022	Adelaide	Shenzhen	6	15	China	
Feb 2022	Sydney	Montreal	1	25	Canada	

An example output data frame would be as follows:

Country	Passengers_in_average	Passengers_out_average	...
Canada	(5 + 1) / 2 = 3.00	(10 + 25) / 2 = 17.50	
Singapore	(8 + 0) / 2 = 4.00	(12 + 0) / 2 = 6.00	
China	(10 + 6) / 2 = 8.00	(20 + 15) / 2 = 17.50	

Output: a data frame (7 columns) with a row for each unique “Country” value, and the following columns: “Passengers_in_average”, “Passengers_out_average”, “Freight_in_average”, “Freight_out_average”, “Mail_in_average”, “Mail_out_average”.

Marks:

0.25 mark – each column with fully correct values and column name

0.5 mark for the correct country list / 0.25 mark for a partially correct country list

Question 4 (2 Marks)

Description: Using the data frame from Question 1, create a new data frame with a descending (highest to lowest) column of the top 5 “Unique_ForeignPort_Count”. If the counts are the same for a country, use alphabetical order.

“Unique_ForeignPort_Count” is the count where routes with at least 1 passenger exist (I.e. where passengers_out > 0) in a unique month from one unique “AustralianPort” to one unique “Country”, but with more than one “ForeignPort” used in that month by that “AustralianPort”. For example, if the input table had the below data, they would be counted as 5 separate occurrences:

1. Sydney to Dubai, United Arab Emirates in September 2022
2. Sydney to Abu Dhabi, United Arab Emirates in September 2022

1. Sydney to Dubai, United Arab Emirates in August 2022
2. Sydney to Abu Dhabi, United Arab Emirates in August 2022

1. Melbourne to Dubai, United Arab Emirates in August 2022
2. Melbourne to Abu Dhabi, United Arab Emirates in August 2022

1. Melbourne to Auckland, New Zealand in September 2022
2. Auckland to Melbourne, New Zealand in September 2022
3. Christchurch to Melbourne, New Zealand in September 2022
4. Melbourne to Queenstown, New Zealand in September 2022
5. Melbourne to Wellington, New Zealand in September 2022

(note – #3 and #4 could be ignored and the outcome will be the same, because inbound routes are disregarded)

1. Brisbane to Shenzhen, China in November 2021
2. Brisbane to Shanghai, China in November 2021

1. Brisbane to Hong Kong, Hong Kong in November 2021 (Note – this is not counted as there is only one ForeignPort used in Hong Kong in this month)

The output for the above would therefore be:

Country	Unique_ForeignPort_Count
United Arab Emirates	3
China	1
New Zealand	1

Output: a data frame with five rows and two columns for “Country” and “Unique_ForeignPort_Count”.

Marks:

1 mark – “Country” column fully correct

1 mark – “Unique_ForeignPort_Count” column fully correct

0.2 mark penalty per incorrect value or incorrect column name

Question 5 (1 marks)

Description: Using the “seats.csv dataset”, load it in a data frame and add two new columns “Source_City” and “Destination_City” to it:

If “In_Out” is “I”, then this flight is inbound to Australia. The “Source_City” is set to the “International_City” and the “Destination_City” is set to the “Australian_City”.

If “In_Out” is “O”, then this flight is outbound from Australia. The “Source_City” is set to the “Australian_City” and the “Destination_City” is set to the “International_City”.

Output: a data frame with all the columns you have loaded from “seats.csv” and the two new columns. The ordering should be preserved from the original CSV file.

Marks:

0.5 mark for each column fully correct

-0.25 mark penalty each if column values are partially correct, a column name is incorrect, or ordering is incorrect

Question 6 (3 marks)

Context: Airlines invest resources to plan specific routes around passenger numbers and competing airlines. For example, Fiji, Jetstar, Qantas, and Virgin fly from Sydney to Nadi. A new airline operating this route may face greater competition than Sydney to Port Vila, which only has Air Vanuatu operating it. However, passenger trends and numbers may have changed over time and a greater number of travellers to Nadi means profitability should still be explored.

Planes are often used on multi city routes. For example, Jetstar can fly to Tokyo via the Gold Coast through Melbourne or Sydney. You may focus on the Australian_City and International_City, as international flight segments of the same plane are already considered in this data set. For example, MEL > SYD > SGN is considered separately for Australian_City values Melbourne and Sydney based on where passengers board.

Description: Using the data frame from Question 5, create a data frame to consolidate statistics for a new or existing airline to understand routes. Your data frame must only include information relevant for airlines to understand the market and how well origin and destination pairs are being serviced. You should add a 250-word comment with optional assumptions to explain your thinking. You may consider dropping unused columns, combining and/or adding values and/or rows, or focusing on regions as you wish.

Output: a data frame.

Marks:

1 mark – table data provides insights that can be used by an airline to understand the market

1 mark – table data is correctly calculated and aligns with the objectives outlined in the comment

0.5 mark – a convincing comment strictly within the word limit that articulates the solutions objectives and how it provides meaningful insight to a new or existing airline

0.5 mark – table data is concise and does not include irrelevant values, rows, or columns

Note: This data set does not include routes with layovers on two different flight numbers. For example, Emirates Airlines is a popular choice for Sydney to London stopovers through Dubai, but they are on two different flight numbers, and only the Sydney to Dubai stopover is included. However, British Airways maintains the same flight number for Sydney to London with a stopover in Singapore. You may focus only on the datasets in this CSV file.

Question 7 (4 marks)

Context: Seat utilisation refers to how full flights are, and airlines optimise their operations to maximise seat utilisation. This has garnered international attention as many airlines oversell seats (https://time.com/6197994/airlines-overbook-flights-negotiate/) on the probability of no shows. To increase profitability, airlines need statistics to bring their passenger numbers closer to the max seat numbers as much as possible. If passenger numbers / maximum seats is < 100%, the airline is running a loss on empty seats. If passenger numbers / maximum seats is > 100%, the airline is running a loss for accommodation, compensation, etc on rebooking passengers. Airlines need statistics to understand their performance and whether their oversell margins or seat numbers need to be adjusted.

Description: From the “city_pairs.csv” and “seats.csv” datasets, across September 2003 to September 2022, and/or the data frames you have created in previous questions, create a visualisation to understand seat utilisation based on the following metrics:

- “Passengers_In”
- “Passengers_Out”
- “Max_Seats”
- “Port_Region”

Output: an image titled “YOUR_ZID_q7.png”. You will be marked based on how good your plot demonstrates trends over time for seat utilisation around the world. You may optionally include multiple plots within one image. Quality of the visualization(s) such as choice of visualization, clarity, choice of colors, labeling/legends, etc. are to be considered in the marking. Be selective about which data you choose to use to uncover your desired insights, e.g. region specific metrics. Include a 250-word comment with optional assumptions to explain your thinking.

Marks:

2 marks – visualisation(s) provide insights that can be used by an airline to understand seat utilisation

1 mark – data used in visualisation(s) are correctly calculated and align with the objectives outlined in the comment

0.5 mark – a convincing comment strictly within the word limit that articulates the solutions objectives and how it provides meaningful insight to airlines looking to optimise seat utilisation

0.5 mark – visualisation(s) are concise and does not include irrelevant values, rows, or columns

Note: 0 marks will be awarded for this question if the image does not save.