Data Scientist

In collaboration with

Overview

Think about a big decision you've had to make, like picking a high school subject, or buying your first car. Usually when we make these big choices, we want to have as much information (or data) about each option, so that we know what we're investing our time, effort and money into.

When it comes to the big choices we make with our money, like opening a savings account, getting into new investments, running a business, or buying your first home, you'll probably look for a bank which is able to give you the best information, support and advice you can get.

All these big decisions are powered by data, and modern banks such as CommBank can collect millions of pieces of data every single day – such as transaction statements that can tell you about how your money goes in and out of your account. They use this data to help make decisions and solve problems, as well as helping to support our customers to make better financial decisions.

To handle all of this important data, banks have teams of dedicated Data Scientists, whose job it is to collect, clean, analyse and visualise this data to help provide better insights and advice for customers.

A Day in the Life

A big part of a Data Scientists everyday job is to collect, clean, analyse and visualise organisation's data – at CommBank this includes financial and customer data to help provide better insights and advice. Let's explore what all of these terms mean in the role of a Data Scientist.

But what do all these terms mean when it comes to treating data?

Collecting data:

Data Scientists use specially designed software that 'pulls' the data from all the different sources (such as mobile banking, transaction histories and statements) and brings it together into one resource.

Cleaning and organising data:

This is one of the most important parts of a data scientist's job. After all, poor quality data, such as data with errors or duplicates could lead to you making incorrect decisions down the line. Data cleaning is the process of making sure any glitches or errors are taken care of using special programs and processes.

Analysing data:

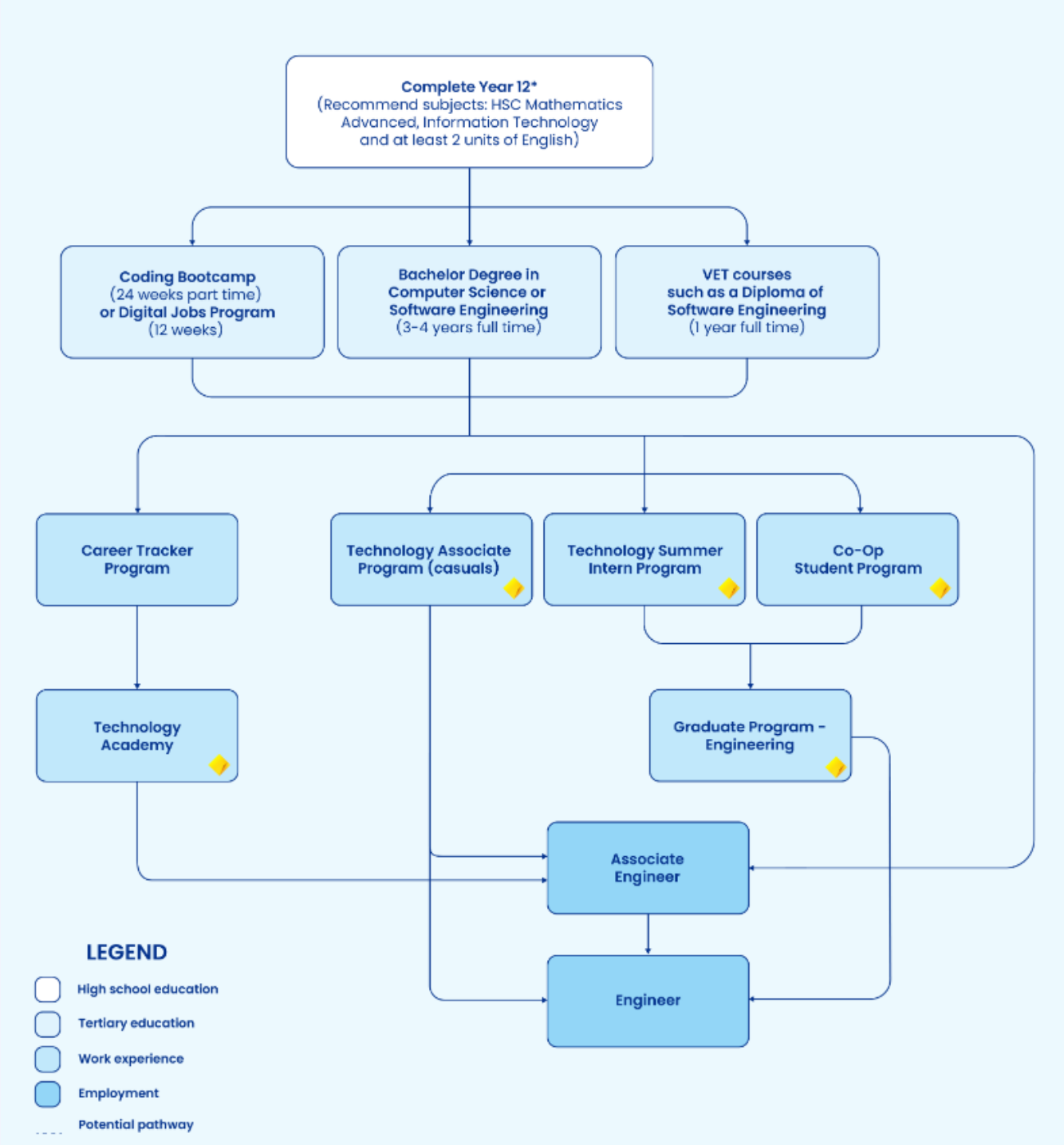
Data scientists at banks use all of this data to create a 'model', or a profile of an individual account, a demographic (think a category of people such as women aged 18-22), an industry, or even an entire country's economy. This helps them figure out important insights and predictions that can improve the financial wellbeing of their customers and communities.

These models can be assisted using cutting-edge technologies such as machine learning and AI to help see things humans would have a hard time detecting, like rapid fraud detection.

Visualising data:

As they say, a picture speaks a thousand words – but data visualisation can tell an even bigger story. When a data scientist receives their data, it's most likely going to be in a big list or spreadsheet that is difficult to read and understand, but by visualising data, you can detect patterns you would have a hard time seeing otherwise, which can help you understand and communicate ideas more effectively.

Pathways


Complete Year 12*
(Recommend subjects: HSC Mathematics Advanced, Information Technology and at least 2 units of English)
Coding Bootcamp
(24 weeks part time)
or Digital Jobs Program
(12 weeks)
Bachelor Degree in Computer Science or Software Engineering
(3-4 years full time)
VET courses such as a Diploma of Software Engineering
(1 year full time)
Career Tracker Program
Technology Associate Program (casuals)
Technology Summer Intern Program
Co-Op Student Program
Technology Academy
Graduate Program - Engineering
Associate Engineer
Engineer
LEGEND
High school education
Tertiary education
Work experience
Employment
Potential pathway

Data Scientist

SCROLL DOWN TO
GET STARTED

ACTIVITY 1

year 13

Commonwealth Bank

Cleaning the Data

Data Science can be very useful to solve many problems. At big organisations like CommBank, a team of Data Scientists is continuously working on finding a solution to the increased number of frauds and scams affecting customers.

In the next 3 activities, you'll assume the role of a Data Scientist working at CommBank, within the Financial Crime Team. Financial crime can involve criminal activity where money and/or property is being stolen, hidden, or used for illegitimate purposes. The main types of financial crime include:

- Fraud
- Money laundering
- Terrorist financing
- Bribery and corruption

It's important to recognise that working for a financial institution, you have the ability and responsibility to help in the reduction and prevention of financial crime. Being part of the Financial Crime team at CommBank, a part of your job is to identify potentially fraudulent transactions from a single customer's transaction history.

A critical part of any data science project is obtaining relevant and reliable data, which can then be used for analysis and modelling.

Our Senior Financial Crime Team members have used Machine Learning and Artificial Intelligence models, which have flagged a single customer as potentially fraudulent. You have data on this customer's financial transactions for a month, which you'll need to analyse and determine whether there is evidence proving there's something suspicious about the activity.

In a real-life scenario where sufficient evidence is found, our frontline teams will take relevant actions to protect the customer's accounts and get in touch with them to discuss the details.

Activity

A commonly agreed principle that is known amongst the data science community is that a model can only perform as well as the data it's given. A very strong model that is run on poor data will almost always give poor results.

This is why it's crucial to make sure that the data is clean before it is used. Now what does it mean when we say that the data must be "clean"?

Data is considered to be unclean when:

There are blank or unrecorded values for necessary fields

As an example, it's quite important that we know the amount of a transaction to be able to conduct an appropriate analysis on that transaction.

There are inconsistent formats within the same field

As an example, if some rows of transactions contain a date in the format DD/MM/YYYY, whereas others contain the date in the format MM/DD/YYYY, it may cause unexpected issues when creating models, analysing the data, or creating diagrams of the data.

Certain data is stored in the wrong format

As an example, let us say that all data in our spreadsheet must be either a date, number, or text. If we have a unique field called "Transaction ID", which we use to identify each transaction, and this field has the format "12345678", this should be stored as a text instead of a number, since the Transaction ID does not represent a numeric quantity but instead acts like a label for the transaction.

For this task, view the table in the next section and see if you can find the cells that require correction.

Transaction Reference	Date	Day	Description	Debit/ Credit	Amount	Method
618125	1/7/2023	Saturday	Woolworths	Debit	$103.12	Credit Card
309851	3/7/2023	Monday	Sydney Water	Debit	$209.00	BPAY
211759	3/7/2023	Monday	Energy Australia	Debit	$312.00	BPAY
567668	4/7/2023	Tuesday	Starbucks	Debit	$6.99	Credit Card
946171	5/7/2023	Wednesday	7-Eleven	Debit	$120.00	Credit Card
978389	6/7/2023	Thursday	Starbucks	Debit	$6.99	Credit Card
158634	7/7/2023	Friday	UBER Eats	Debit	$35.00	Credit Card
503948	8/7/2023	Saturday	Woolworths	Debit	$107.44	Credit Card
782522	10/7/2023	Monday	Optus	Debit	$45.00	Credit Card
939042	11/7/2023	Tuesday	Starbucks	Debit	$6.99	Credit Card
385230	12/7/2023	Wednesday	Salary	Credit	$2500.00	Bank Transfer
192517	13/7/2023	Thursday	Starbucks	Debit	$6.99	Credit Card
483632	13/7/2023	Thursday	UBER Eats	Debit	$32.50	Credit Card
541553	14/7/2023	Friday	Dinner	Debit	$74.99	Credit Card
427818	15/7/2023	Saturday	Netflix	Debit	$10.99	Credit Card
555610	15/7/2023	Saturday	Woolworths	Debit	$97.86	Credit Card
535139	19/7/2023	Wednesday	7-Eleven	Debit	$120.00	Credit Card
437075	19/7/2023	Wednesday		Debit	$6.99	Credit Card
755572	20/7/2023	Thursday	Bunings	Debit	$350.00	PAYID
681353	20/7/2023	Thursday	Christmas Gift	Debit	$499.00	PAYID

770672	21/7/2023	Friday	Dinner	Debit	$999.00	PAYID
702581	22/7/2023	Saturday	Woolworths	Debit	$89.00	Credit Card
624035	22/7/2023	Saturday	Telstra	Debit	$250.00	Credit Card
449468	22/7/2023	Saturday	Telstra	Debit	$250.00	Credit Card
919306	22/7/2023	Saturday	Telstra	Debit	$250.00	Credit Card
392433	07/25/2023	Tuesday	Starbucks	Debit	$6.99	Credit Card
720897	26/7/2023	Wednesday	Salary	Credit	$2500.00	Bank Transfer
501684	28/7/2023	Friday	Hoyts	Debit	$50.00	Credit Card
610200	28/7/2023	Friday	Starbucks	Debit	6.99	Credit Card
337811	29/7/2023	Saturday	Woolworths	Debit	$112.76	Credit Card

Q1. Enter the Transaction ID of the row which has no description listed:

- ◯ 309851
- ◯ 437075
- ◯ 939042
- ◯ 555610

Q2. Enter the Transaction ID of the row which has an incorrect date format:

- 702581
- 427818
- 211759
- 392433

Q3. Enter the Transaction ID of the row which has the incorrect format for amount:

- 782522
- 720897
- 555610
- 610200

Q4. Using the other prices in the table as a reference, which vendor do you think transaction reference 437075 is from?

- ○ Woolworths
- ○ Optus
- ○ Starbucks
- ○ Energy Australia

All done!

As you can imagine, data is highly sensitive and is something that must be handled with confidentiality and care. Take CommBank for example, with a customer base of 17 million customers, and each of those customers' sensitive data stored within their databases, it is crucial that:

1. The data is kept secure, and sufficient protocols and measures are kept in place to ensure the data does not end up in the wrong hands.
2. The data is used ethically, and it is not used in a way that customers would find inappropriate.
3. The data is relevant, and that any data that does not assist the Bank in improving how customers are serviced is discarded appropriately.

Within the world of data science, there are a lot of ethical considerations to keep in mind. Data scientists often work with highly personal information, so precautions must be taken to ensure that the data is treated with the necessary care. Once data is confirmed to be clean, data scientists can then begin to analyse the data, which we will look at in the next activity. Check your answers below and then jump into activity 2!

Answers

1. b) 437075
2. d) 392433
3. d) 610200
4. c) Starbucks

SCROLL DOWN TO
GET STARTED

Data Scientist

ACTIVITY 2

year 13

Commonwealth Bank

Categorising the Data

An important component of a Data Scientist's work is to be able to make sense of the data. There are numerous ways this can be achieved. One way is to categorise data into groups that can provide additional clarity.

How could machine learning help?

Typically, with an organisation as large as CommBank, it's not feasible for our Data Science teams to manually go through transactions and categorise them. This is where concepts like Machine Learning are incredibly useful and instrumental.

Machine Learning is a type of Artificial Intelligence that focuses on using data, algorithms and frameworks, to allow machines to obtain insights and analytics.

Some of the ways in which Machine Learning could be implemented in this scenario include:

Analysing Transaction Descriptions

This is to determine what type of transaction it is (e.g., transactions with a description of a clothing brand would most likely represent transactions that involve shopping and/or clothing).

Analysing Transaction Costs

This is so the transaction can be compared to the typical amount that is spent on certain items (e.g. purchases from fast-food stores are usually lower than purchases from a restaurant).

Analysing Transaction Frequencies

This is to determine if this is a regular transaction from this customer, or potentially an unnatural and suspicious transaction (e.g., if a customer typically spends $100 per month at restaurants, but has a $900 transaction occurring, it may raise flags as it does not align with the standard behaviour of this customer).

Activity

For this task, select which purchases from the bank statement fit into each of the categories in the questions below:

Q1. Which of the following purchases would fit into the ENTERTAINMENT category?

- [ ] Woolworths
- [ ] Sydney Water
- [ ] Energy Australia
- [ ] Starbucks
- [ ] 7-Eleven (Fuel)
- [ ] UBER Eats
- [ ] Optus
- [ ] Dinner
- [ ] Netflix
- [ ] Bunnings
- [ ] Telstra
- [ ] Hoyts

Q2. Which of the following purchases would fit into the UTILITY BILLS category?

- [ ] Woolworths
- [ ] Sydney Water
- [ ] Energy Australia
- [ ] Starbucks
- [ ] 7-Eleven (Fuel)
- [ ] UBER Eats
- [ ] Optus
- [ ] Dinner
- [ ] Netflix
- [ ] Bunnings
- [ ] Telstra
- [ ] Hoyts

Q3. Which of the following purchases would fit into the FOOD category?

- [ ] Woolworths
- [ ] Sydney Water
- [ ] Energy Australia
- [ ] Starbucks
- [ ] 7-Eleven (Fuel)
- [ ] UBER Eats
- [ ] Optus
- [ ] Dinner
- [ ] Netflix
- [ ] Bunnings
- [ ] Telstra
- [ ] Hoyts

Q4. Which of the following purchases would fit into the TRAVEL category?

- [ ] Woolworths
- [ ] Sydney Water
- [ ] Energy Australia
- [ ] Starbucks
- [ ] 7-Eleven (Fuel)
- [ ] UBER Eats
- [ ] Optus
- [ ] Dinner
- [ ] Netflix
- [ ] Bunnings
- [ ] Telstra
- [ ] Hoyts

Q5. Which of the following purchases would fit into the OTHER category?

- [ ] Woolworths
- [ ] Sydney Water
- [ ] Energy Australia
- [ ] Starbucks
- [ ] 7-Eleven (Fuel)
- [ ] UBER Eats
- [ ] Optus
- [ ] Dinner
- [ ] Netflix
- [ ] Bunnings
- [ ] Telstra
- [ ] Hoyts

All done!

Machine Learning capitalises on all these factors, and many more, to be able to formulate insights into data on a large scale. Analysing 17 million individual customers' transaction behaviour manually may not be possible, but Machine Learning allows the task to be completed in a much more regular and feasible manner.

Once the data has been analysed, data scientists can get into the grit of their work – using that data to solve a problem. Move on to Activity 3 to find out how they do this!

Answers

1. Netflix, Hoyts
2. Sydney Water, Energy Australia, Optus, Telstra
3. Woolworths, Starbucks, UBER Eats, Dinner
4. 7-Eleven (Fuel)
5. Bunnings, Christmas Gift

Data Scientist

year 13

Commonwealth Bank

ACTIVITY 3

SCROLL DOWN TO GET STARTED

Identifying Anomalies in the Data

Once the data is obtained, cleaned, and basic analysis is completed, the next step is to apply our understanding of the data to the problem we are trying to solve. In this case, the objective is to determine if this customer has any suspicious transactions on their account.

A simple method that you can use to identify suspicious transactions is to look for outliers/anomalies, or unnatural transaction behaviour.
Simply put, outliers/anomalies are data points which do not follow the same pattern as the majority of the data.

Activity

For this task, using the table below, look at the descriptions, amounts and frequencies of the transactions in the table below to see if you can find the most suspicious transactions.

Then, answer the questions below the table. Remember that:

- Some transactions may be suspicious due to the type and frequency.
- Some transactions may be suspicious due to having an unnatural description.
- Some transactions may be suspicious due to having an unnatural amount.

Transaction Reference	Date	Day	Description	Debit/ Credit	Amount	Method
618125	1/7/2023	Saturday	Woolworths	Debit	$103.12	Credit Card
309851	3/7/2023	Monday	Sydney Water	Debit	$209.00	BPAY
211759	3/7/2023	Monday	Energy Australia	Debit	$312.00	BPAY
567668	4/7/2023	Tuesday	Starbucks	Debit	$6.99	Credit Card
946171	5/7/2023	Wednesday	7-Eleven	Debit	$120.00	Credit Card
978389	6/7/2023	Thursday	Starbucks	Debit	$6.99	Credit Card
158634	7/7/2023	Friday	UBER Eats	Debit	$35.00	Credit Card

503948	8/7/2023	Saturday	Woolworths	Debit	$107.44	Credit Card
782522	10/7/2023	Monday	Optus	Debit	$45.00	Credit Card
939042	11/7/2023	Tuesday	Starbucks	Debit	$6.99	Credit Card
385230	12/7/2023	Wednesday	Salary	Credit	$2500.00	Bank Transfer
192517	13/7/2023	Thursday	Starbucks	Debit	$6.99	Credit Card
483632	13/7/2023	Thursday	UBER Eats	Debit	$32.50	Credit Card
541553	14/7/2023	Friday	Dinner	Debit	$74.99	Credit Card
427818	15/7/2023	Saturday	Netflix	Debit	$10.99	Credit Card
555610	15/7/2023	Saturday	Woolworths	Debit	$97.86	Credit Card
535139	19/7/2023	Wednesday	7-Eleven	Debit	$120.00	Credit Card
437075	19/7/2023	Wednesday		Debit	$6.99	Credit Card
755572	20/7/2023	Thursday	Bunings	Debit	$350.00	PAYID
681353	20/7/2023	Thursday	Christmas Gift	Debit	$499.00	PAYID
770672	21/7/2023	Friday	Dinner	Debit	$999.00	PAYID
702581	22/7/2023	Saturday	Woolworths	Debit	$89.00	Credit Card
624035	22/7/2023	Saturday	Telstra	Debit	$250.00	Credit Card
449468	22/7/2023	Saturday	Telstra	Debit	$250.00	Credit Card
919306	22/7/2023	Saturday	Telstra	Debit	$250.00	Credit Card
392433	07/25/2023	Tuesday	Starbucks	Debit	$6.99	Credit Card
720897	26/7/2023	Wednesday	Salary	Credit	$2500.00	Bank Transfer
501684	28/7/2023	Friday	Hoyts	Debit	$50.00	Credit Card
610200	28/7/2023	Friday	Starbucks	Debit	6.99	Credit Card
337811	29/7/2023	Saturday	Woolworths	Debit	$112.76	Credit Card

Q1. Enter the Transaction ID of the first suspicious transaction (highest in the table):

- ○ 755572
- ○ 385230
- ○ 618125
- ○ 211759

Q2. Enter the Transaction ID of the second suspicious transaction:

- ○ 946171
- ○ 158634
- ○ 681353
- ○ 192517

Q3. Enter the Transaction ID of the third suspicious transaction:

- ○ 541553
- ○ 483632
- ○ 770672
- ○ 385230

Q4. Enter the Transaction ID of the fourth suspicious transaction:

- ○ 624035
- ○ 535139
- ○ 702581
- ○ 392433

Q5. Enter the Transaction ID of the fifth suspicious transaction:

- ○ 337811
- ○ 501684
- ○ 702581
- ○ 449468

Q6. Enter the Transaction ID of the sixth suspicious transaction:

- ○ 483632
- ○ 919306
- ○ 610200
- ○ 501684

All done!

Well done on completing this activity! You've just gained an insight into some key elements of what data science involves.

The key takeaways from these activities are:

1. It's important to investigate the quality of your data before carrying out an analysis.
2. It's useful to break down your data into ways in which it is easier to understand.
3. It's crucial to investigate and analyse data whilst using your logical reasoning and thinking skills.

Answers

1. a) Transaction ID 755572 may be suspicious since Bunnings is spelt incorrectly, the amount of the transaction is slightly unnatural, and it is not usual for a customer to pay Bunnings using PayID instead of a Debit/Credit Card.

2. c) Transaction ID 681353 may be suspicious since a large amount is being transferred, and the description indicates it is for a Christmas gift, despite the date being in July. You would expect a Christmas gift to be given closer to December.

3. c) Transaction ID 770672 may be suspicious since a large amount is being transferred for the purpose of a dinner. Previous transactions for this customer's salary and dinner indicate that the amount they have spent on this dinner seems unreasonable.

4. a) Transaction ID 624035 may be suspicious since the amount indicates that the customer is possibly purchasing some sort of SIM Card package, however the customer purchasing this three times within the same day seems potentially suspicious.

5. Transaction ID 449468 may be suspicious since the amount indicates that the customer is possibly purchasing some sort of SIM Card package, however the customer purchasing this three times within the same day seems potentially suspicious.

6. b) Transaction ID 919306 may be suspicious since the amount indicates that the customer is possibly purchasing some sort of SIM Card package, however, the customer purchasing this three times within the same day seems potentially suspicious.