Data Science WIL final Project

Jermaine Varnicker

Student number: 8663

CTU Training Solutions

Auckland Park

Instructor's Name: Newton Nkeng

March 2021

Summary

The focus of this report is to showcase some of the basic functions of a data scientist and through scenario-based problems how Data science can solve business problems in the real world. The report is made up of 3 sections and 3 scenarios, each scenario is targeted at highlighting a particular skillset that is essential to the Data science function.
To give a brief overview of how the report is set up:

Focus area for each section:
Section 1: Statistics
Section 2: Power BI: reporting and app deployment
Section 3: Programming in python and Manipulating Data in T-SQL

Contents

Introduction

The objective of this report is to demonstrate my ability to apply the knowledge obtained through the FNB data science learnership, on an industry level. Each section of this report will cover a different scenario which will bring into focus a specific skill/tool or function that is essential to the role of a data scientist. By the end of this report my hope is that the solutions I have provided for the different scenario-based questions will prove to the reader my ability to perform the tasks of a data scientist within a working environment.

Section 1:

Focus Area Statistics

Statistical research gives businesses the information they need to make informed decisions in uncertain circumstances. When stakeholders analyze statistical research in business, they determine how to proceed in areas including auditing, financial analysis and marketing research. To further elaborate on the former, I will demonstrate how statistics can aid in making more informed business decisions.

The scenario in this section has 4 problems to be solved. Each question is followed by the solution with a brief explanation of the steps performed.

Scenario 1:

Number of questions: 4

Suppose that you are deciding between two alternative investments. Investment A is a mutual fund whose portfolio consists of a property exchange-traded fund (ETF), which are both accessible via stock exchanges. Investment B consists of shares of a growth stock. You estimate the returns (per R1, 000 investment) for each investment alternative under three economic condition events (recession, stable economy, and growing economy). The three economic conditions for the two alternative investments are calculated as follows:

Investment A

- Recession: -10% per R1, 000 investment
- Stable economy: +10% per R1, 000 investment
- Growing economy: +20% per R1, 000 investment

Investment B

- Recession: -20% per R1, 000 investment
- Stable economy: +35% per R1, 000 investment
- Growing economy: +45% per R1, 000 investment

The subjective probability of the occurrence of each economic condition, in both investments, is presented as follows:

- Recession: 20%
- Stable economy: 50%
- Growing economy: 30%

Questions and Solutions for Section 1:

Question 1:

Calculate the expected return for the two investments for the next 10 years to come and create a line chart for both investments.

Solution to Question 1:

The first step to solving the problem was to calculate the mean based on the provided economic conditions the calculations below demonstrate how that is done:

mean/expected value for investment A

$\mu = (-0.1)(0.2) + (0.1)(0.5) + (0.2)(0.3) = 0.09 \; or \; 9\%$

$\mu = R90$

mean/expected value for investment B

$\mu = (-0.2)(0.2) + (0.35)(0.5) + (0.45)(0.3) = 0.27 \; or \; 27\%$

$\mu = R270$

The next step was to calculate the expected return for the two investments over the next 10 years. This was calculated using the expected value formula: $\boldsymbol{E(X) = P(x) * X}$

NB: to avoid repeating the same calculations for each year, the projections have been placed in table A and table B respectively as shown below:

Investment A over 10 years

Table A: Expected return on investment A

Equation used:	$\boldsymbol{E(X) = P(x) * X} = E(X) = 0.09$ * year number									
(X)Year	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
EV	0.09	0.18	0.27	0.36	0.45	0.54	0.63	0.72	0.81	0.9
Per R1000	R90	R180	R270	R360	R450	R540	R630	R720	R810	R900

Investment B over 10 years

Table B:Expected return on investment B

Equation used:	$\boldsymbol{E(X) = P(x) * X} = E(X) = 0.27$ * year number									
(X)Year	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
EV	0.27	0.54	0.81	1.08	1.35	1.62	1.89	2.16	2.43	2.7
Per 1000	R270	R540	R810	R1080	R1350	R1620	R1890	R2160	R2430	R2700

In the final step the line chart was created to visually display the expected value as it is represented in table A and B.

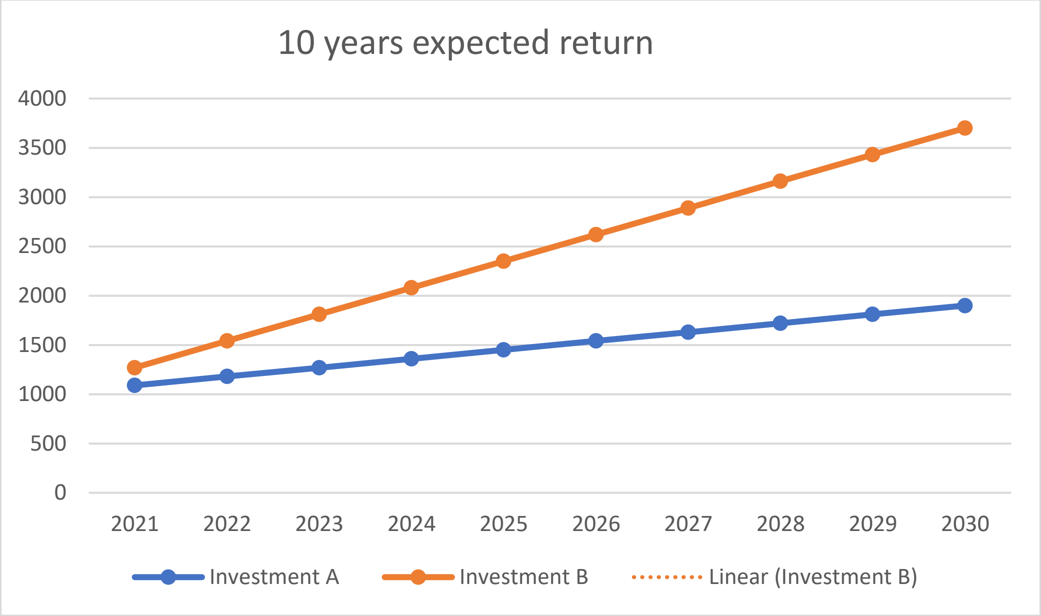


Figure 1. line chart for both investments

Question 2:

Computing the Variance and Standard Deviation for Property exchange-traded fund (A)

Solution to Question 2:

To calculate the variance the following logic was applied: Variance = Sum of [(Each value-mean)^2 x probability]

Variance for investment A

$$\sigma^2 = (-100 - 0.09)^2 \times (0.2) + (100 - 0.09)^2 \times (0.5) + (200 - 0.09)^2 \times (0.3)$$

$$\sigma^2 = 18983.80$$

The Standard Deviation is just the square root of Variance. see below solution:

Standard deviation for investment A

$$SD = \sqrt{18983.80}$$

$$SD = 137.78$$

Question 3:
Computing the Variance and Standard Deviation for Growth Stock (B)

Solution to Question 3:

NB: Questions 3 follows the same logical steps as Q2 so to avoid being redundant no explanation will be provided.

<u>Variance for investment B</u>

$\sigma^2 = (-200 - 0.27)^2 \times (0.2) + (350 - 0.27)^2 \times (0.5) + (450 - 0.27)^2 \times (0.3)$
$\sigma^2 = 129854.27$

<u>Standard deviation for investment B</u>

$SD = \sqrt{129854.27}$
$SD = \ 360.35$

Question 4:
Identify which investment present a better opportunity based on the mean return and the standard deviation.

Solution to Question 4:
Based on the standard deviation and mean of Investment A when compare to Investment B we can deduce that even though investment B has a higher mean value it also has a higher standard deviation and what this means is that the expected values/data will move farther from its mean value under different economic conditions thus making it a volatile or risky investment.

Investment A having a lower mean value tells us that its overall returns will be less than that of investment B but because investment A has a lower standard deviation we can deduce that it will perform more stable under different economic conditions as that expected values/data point will not move very far from its mean value making it the less risky option and the better investment choice.

END OF SECTION 1

Section 2

Focus Area: Power BI- reporting

Power BI is one of the most widely used Business intelligence tools on the market today. There are several benefits to a business incorporating BI into the business models. The primary benefit of business intelligence tools is to provide a solution that is suitable for a business. It customizes and offers a solution that is most suitable to the company and is relevant to the business goals. With business intelligence and business analytics, your company can have an account of all the impacts it has had be it good or bad. Data analytics plays a vital role in helping your company devise the best plan possible. Business intelligence systems help companies move towards the right direction and make smart choices from the business data collected and analyzed. The following scenario will demonstrate this.

Please note that in this section the solutions to the questions are supplemented with images as the actual work was done on the power BI platform. At the end of this section, I have provided links to the downloadable reports and data as well as a link to the published app which can be used to see which filters were applied to generate the required outcome of each question.

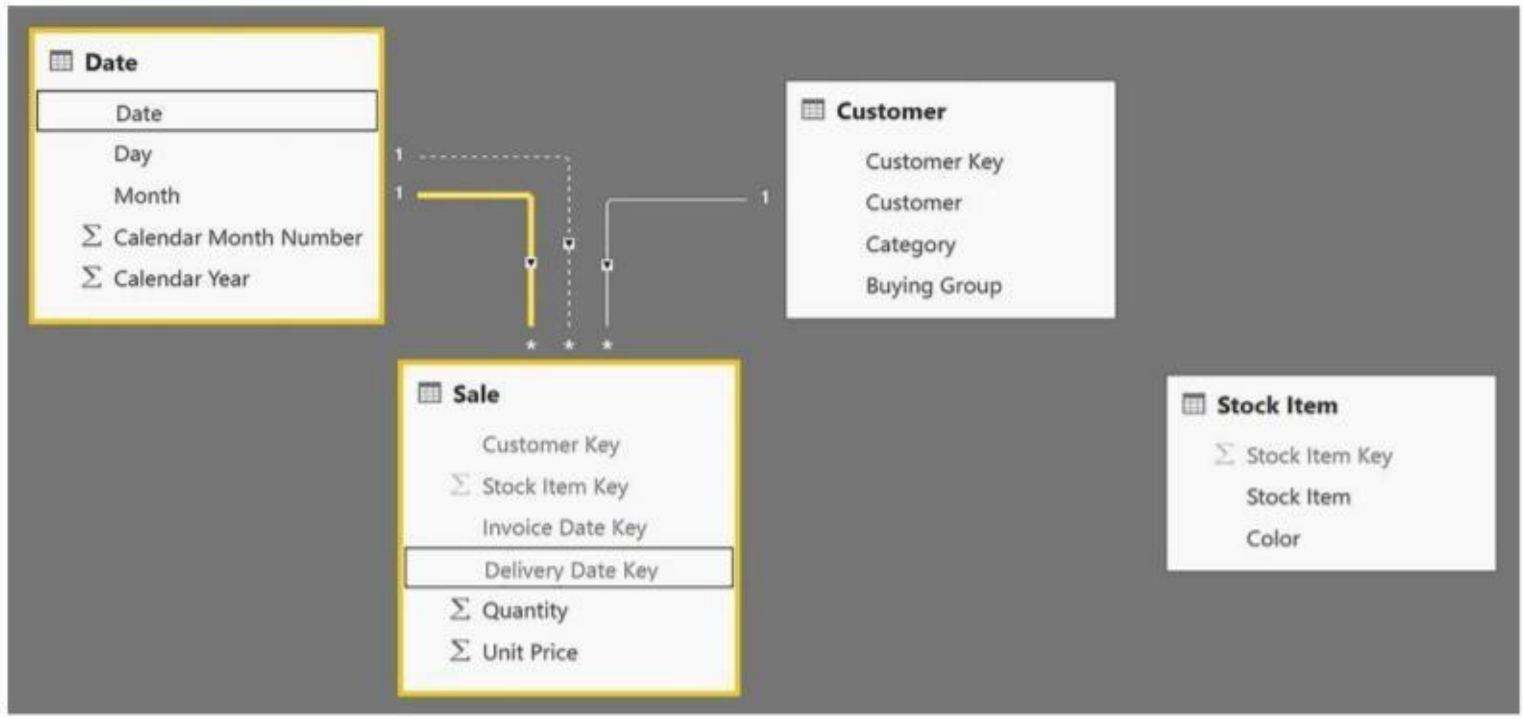
The scenario follows on the next page.

Scenario 2:

Number of questions: 9

You are the BI developer at Wide World Importers responsible for creating Power BI reports. You have already connected to data sources and imported data in Power BI Desktop. Below is presented the data WAREHOUSE DOWNLOAD LINK. https://github.com/Microsoft/sql-server-samples/releases/download/wide-worldimporters-v1.0/WideWorldImportersDW-Full.bak

The data model is presented as follows:

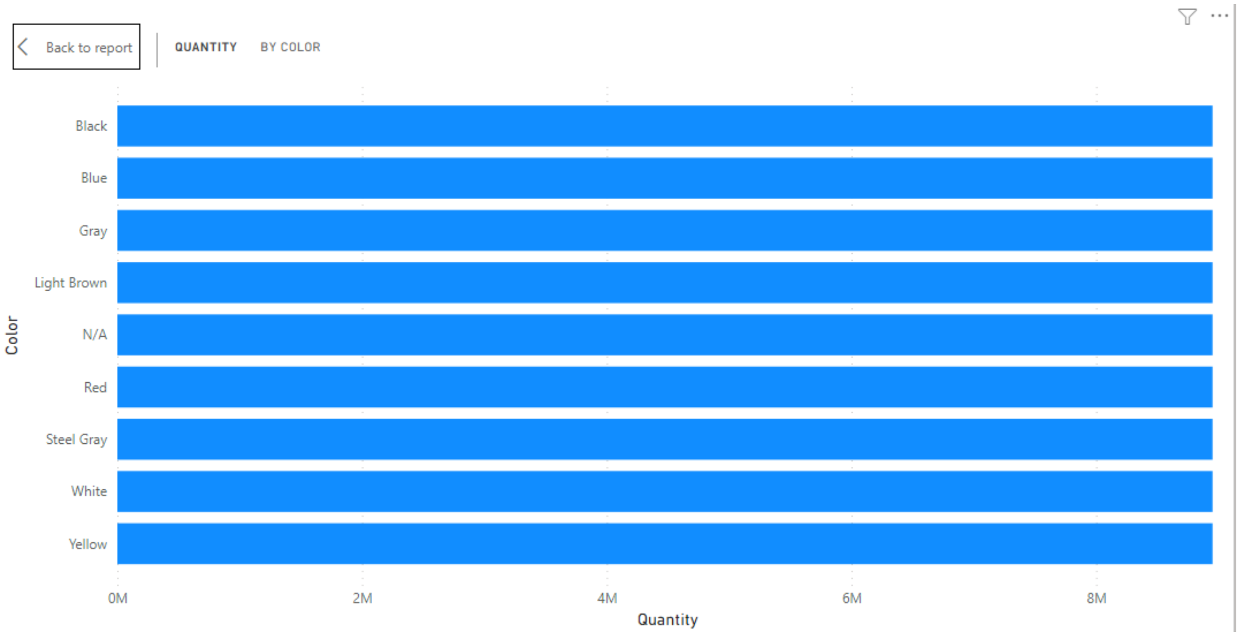


The management requested a report based on historical data available. Based on background information and business requirements, answer the following questions:

Questions and Solutions for Section 2:

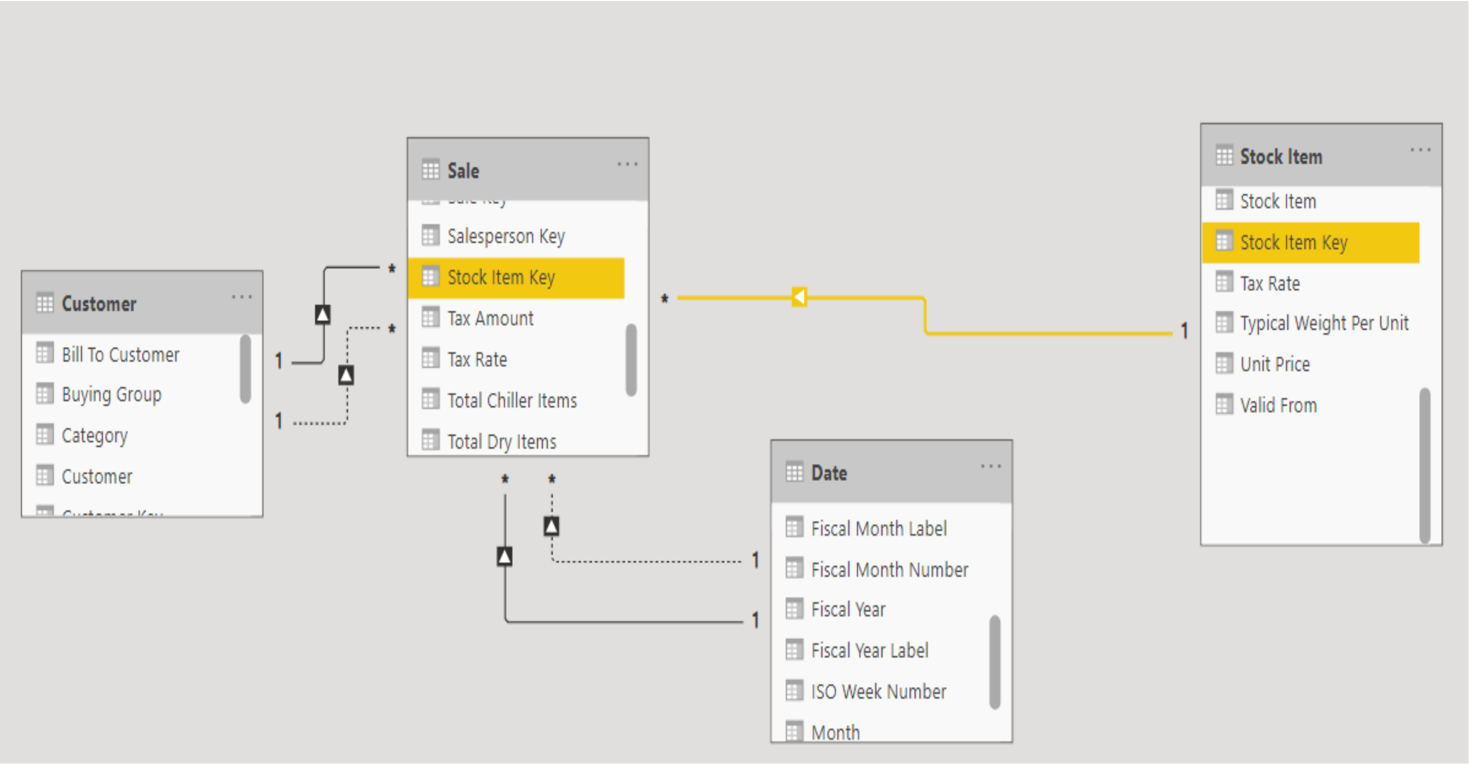
Question 1:

Create a bar graph with Color on the axis and Quantity on the values.

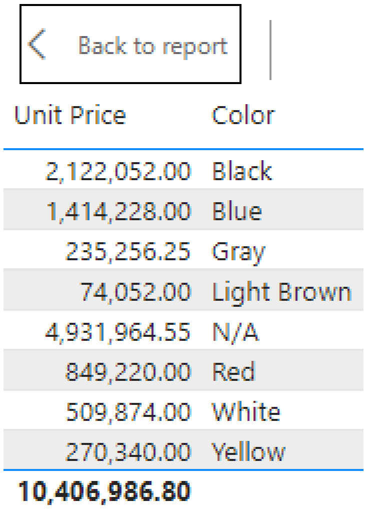


Question 2:

Create an active physical relationship between Sale and Stock Item.

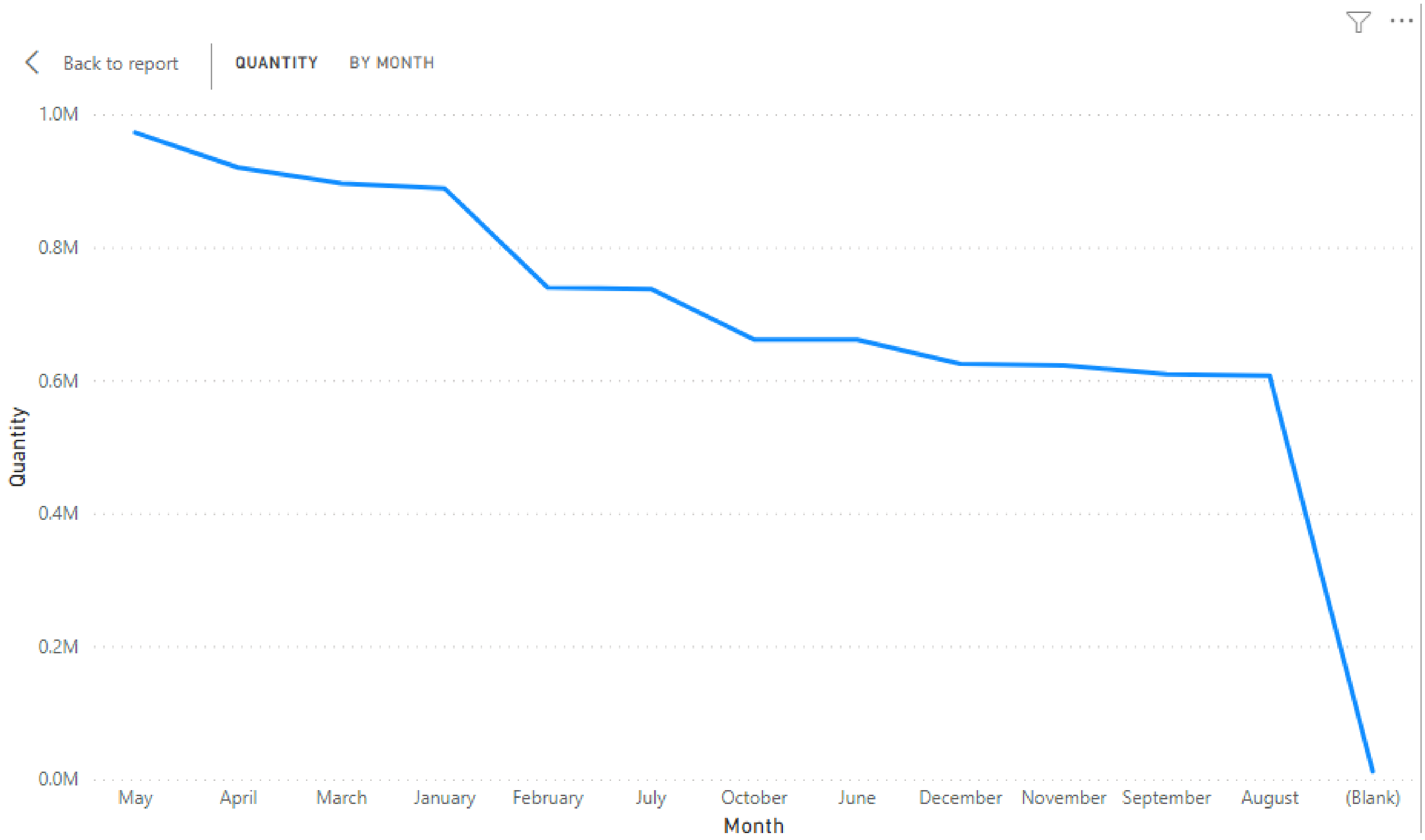


Question 3:
Create a filtered calculated table for each color based on the Sale table.



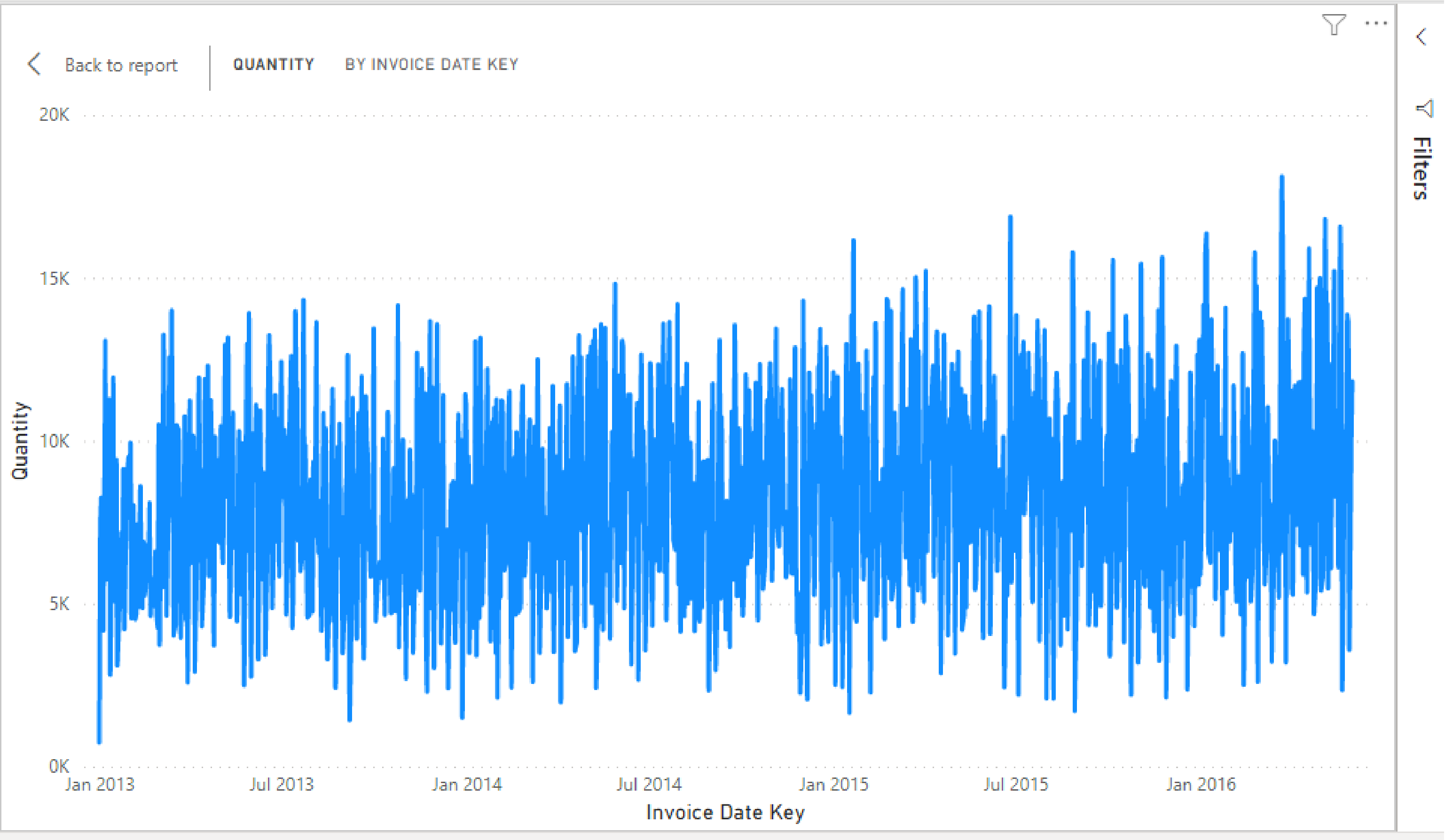
Unit Price	Color
2,122,052.00	Black
1,414,228.00	Blue
235,256.25	Gray
74,052.00	Light Brown
4,931,964.55	N/A
849,220.00	Red
509,874.00	White
270,340.00	Yellow
10,406,986.80	

Question 4:
Create a line chart that shows Quantity by Month.



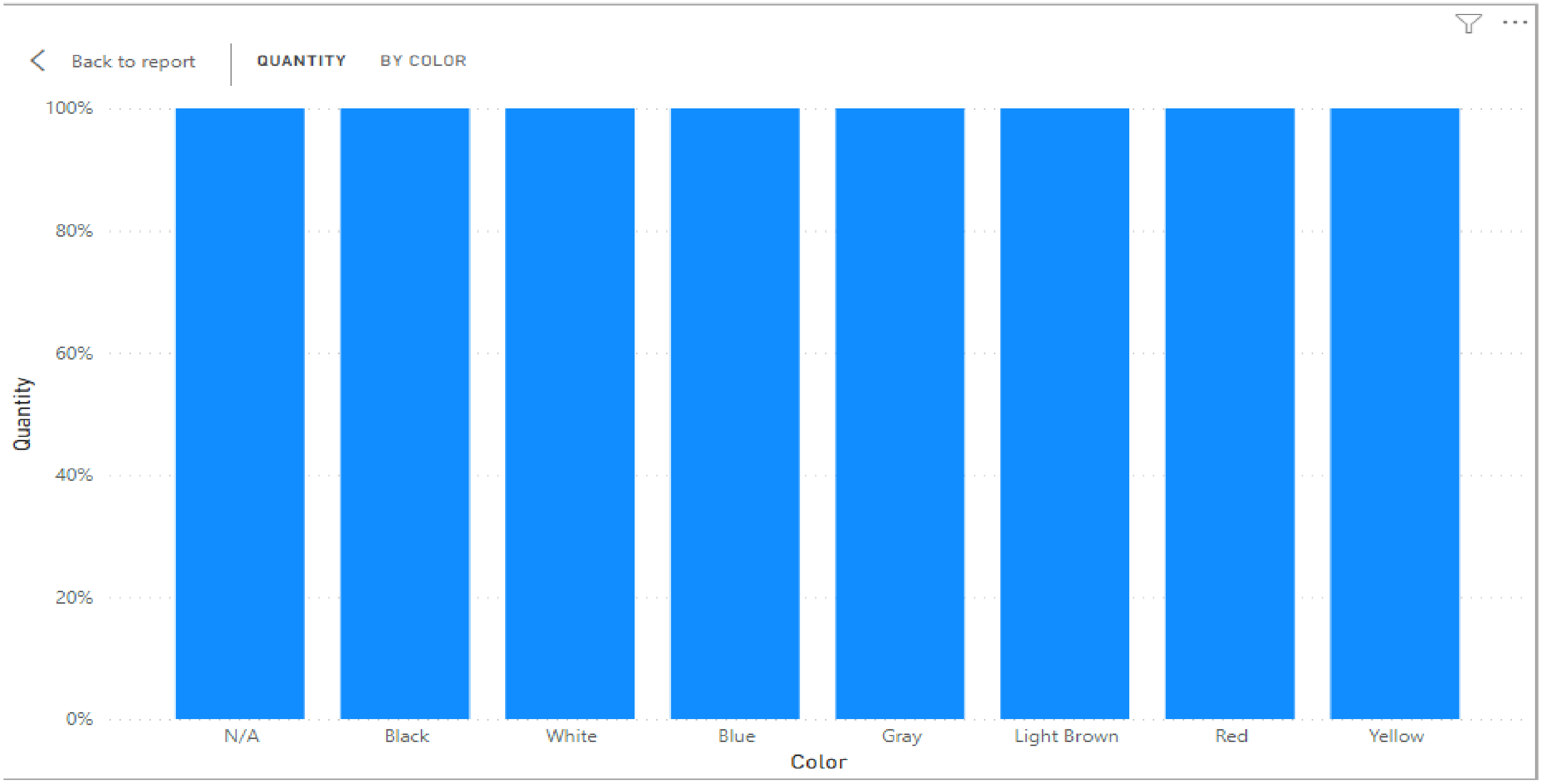
Question 5:

Create a line chart showing Quantity by the invoice date



Question 6:

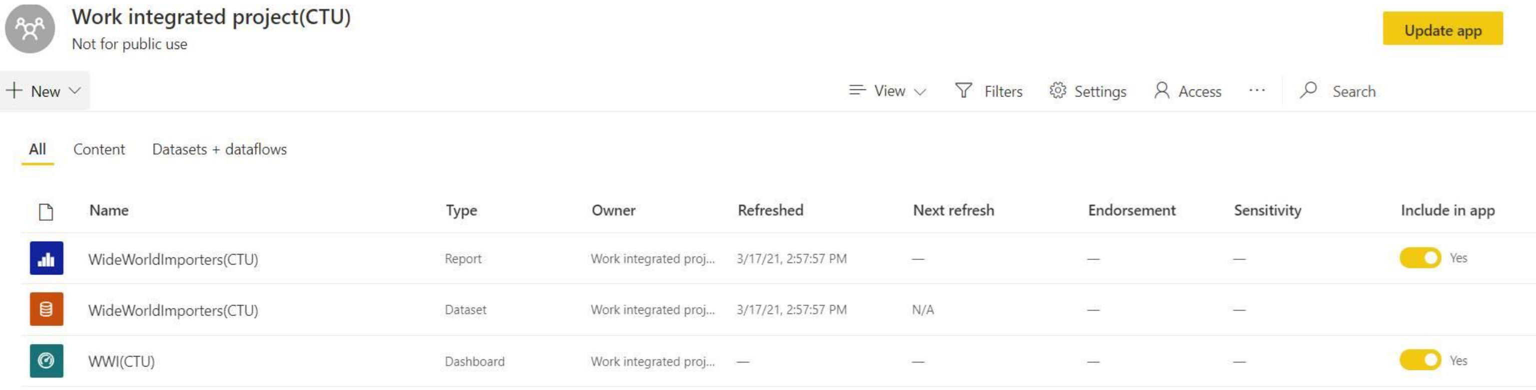
Create a column chart that displays Quantity by Color



Question 7 and 8:

Create and configure an app workspace for the created model, Publish the app

Workspace For app deployment

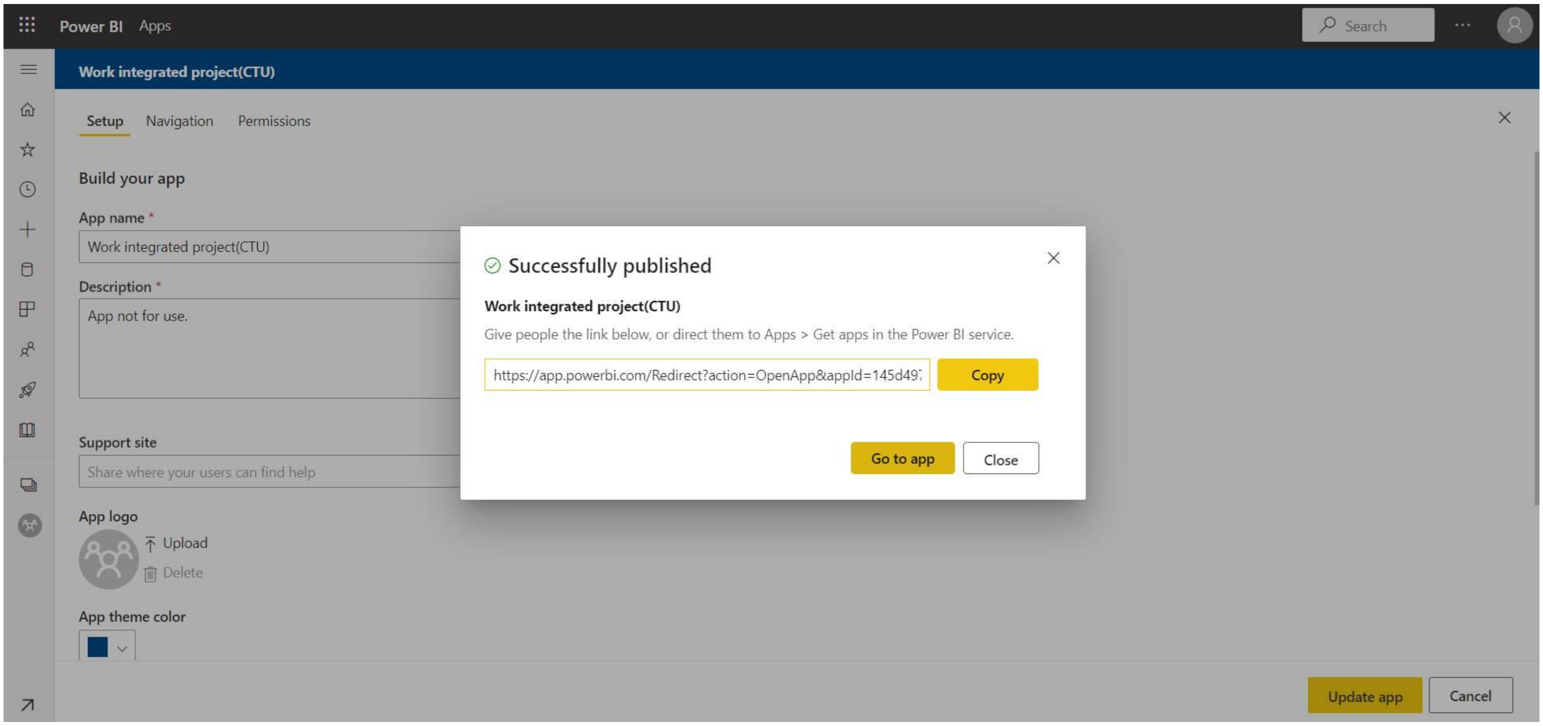


Work integrated project(CTU)

Not for public use

Name	Type	Owner	Refreshed	Next refresh	Endorsement	Sensitivity	Include in app
WideWorldImporters(CTU)	Report	Work integrated proj...	3/17/21, 2:57:57 PM	—	—	—	Yes
WideWorldImporters(CTU)	Dataset	Work integrated proj...	3/17/21, 2:57:57 PM	N/A	—	—	
WWI(CTU)	Dashboard	Work integrated proj...	—	—	—	—	Yes

App Successfully published



App name: Work integrated project(CTU)

Description: App not for use.

Successfully published

Work integrated project(CTU)

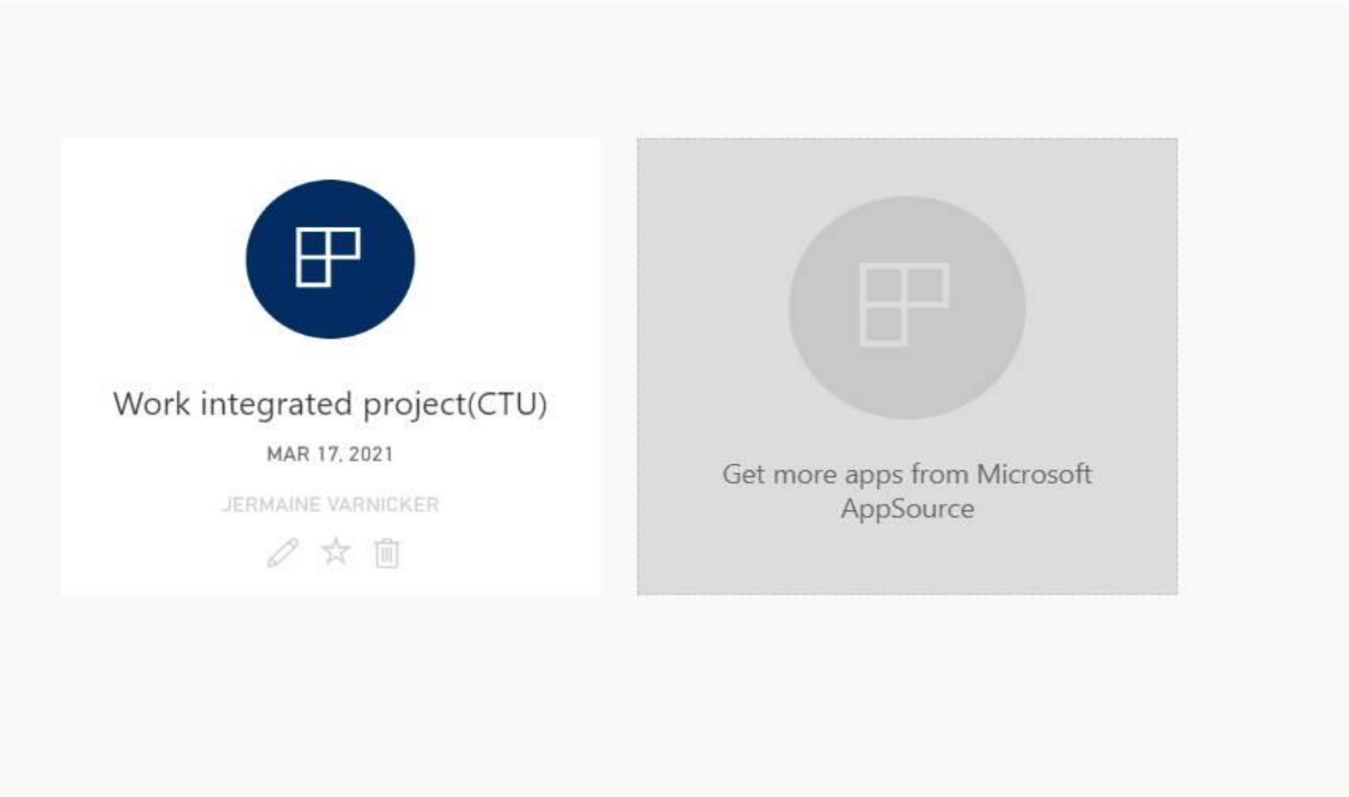
Give people the link below, or direct them to Apps > Get apps in the Power BI service.

https://app.powerbi.com/Redirect?action=OpenApp&appId=145d497

Link to app: https://app.powerbi.com/Redirect?action=OpenApp&appId=145d4970-7cc5-471f-9f61-6959865f9248&ctid=f5ea3467-a1df-4d7e-a894-9a0c66d9b19e

Question 9:

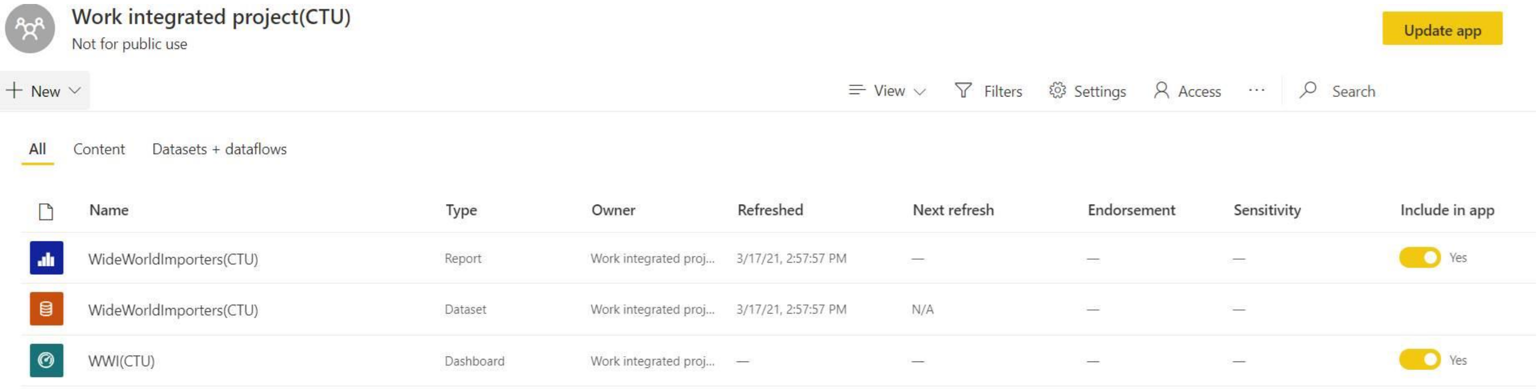
Package dashboards and reports as apps



Work integrated project(CTU)

MAR 17, 2021

JERMAINE VARNICKER



Work integrated project(CTU)

Not for public use

	Name	Type	Owner	Refreshed	Next refresh	Endorsement	Sensitivity	Include in app
	WideWorldImporters(CTU)	Report	Work integrated proj...	3/17/21, 2:57:57 PM	—	—	—	Yes
	WideWorldImporters(CTU)	Dataset	Work integrated proj...	3/17/21, 2:57:57 PM	N/A	—	—	
	WWI(CTU)	Dashboard	Work integrated proj...	—	—	—	—	Yes

Link to download the pbxi file: https://1drv.ms/u/s!AuRvf69NQWuRiHIzIRMFrGJ4dMYC?e=C5nXoG

Link to app: https://app.powerbi.com/Redirect?action=OpenApp&appId=145d4970-7cc5-471f-9f61-6959865f9248&ctid=f5ea3467-a1df-4d7e-a894-9a0c66d9b19e

END OF SECTION 2

Section 3:

Focus Area: Programming in python and Manipulating Data in T-SQL

When it comes to Data Science, Python is a an excellent tool with a whole range of benefits. Since it is open source, it is flexible and continuously improving. In addition python has an array of useful libraries and not to forget that it can be integrated with other languages like Java.

Python is one of the most common programming languages adopted by data scientists and database administrators, the ability to run Python code as T-SQL script enables the machine learning capabilities, directly when it comes to dealing with large amounts of data. As we know Data Science is the study and analysis of data. To analyze the data, we need to extract it from the database and this is where T-SQL comes into the picture.

Knowing how to navigate program data hierarchies, or big data, and query certain datasets alongside knowing how to code algorithms and develop models is invaluable to a data scientist. The scenario along with the solutions in this section aims to demonstrate this.

Please note: I have provided a video link to supplement question 3.3 the video link also serves as a "walk-though" explanation of how each problem in the scenario was solved.

Scenario 3:

Number of questions: 3

You are provided with two csv files (The files are placed on ColCampus under the project) Write a Python program (Use any python IDE of your choice) to join the two given files along the common column. The output file is named "combined_csv. csv".

Perform the following on the combined file:

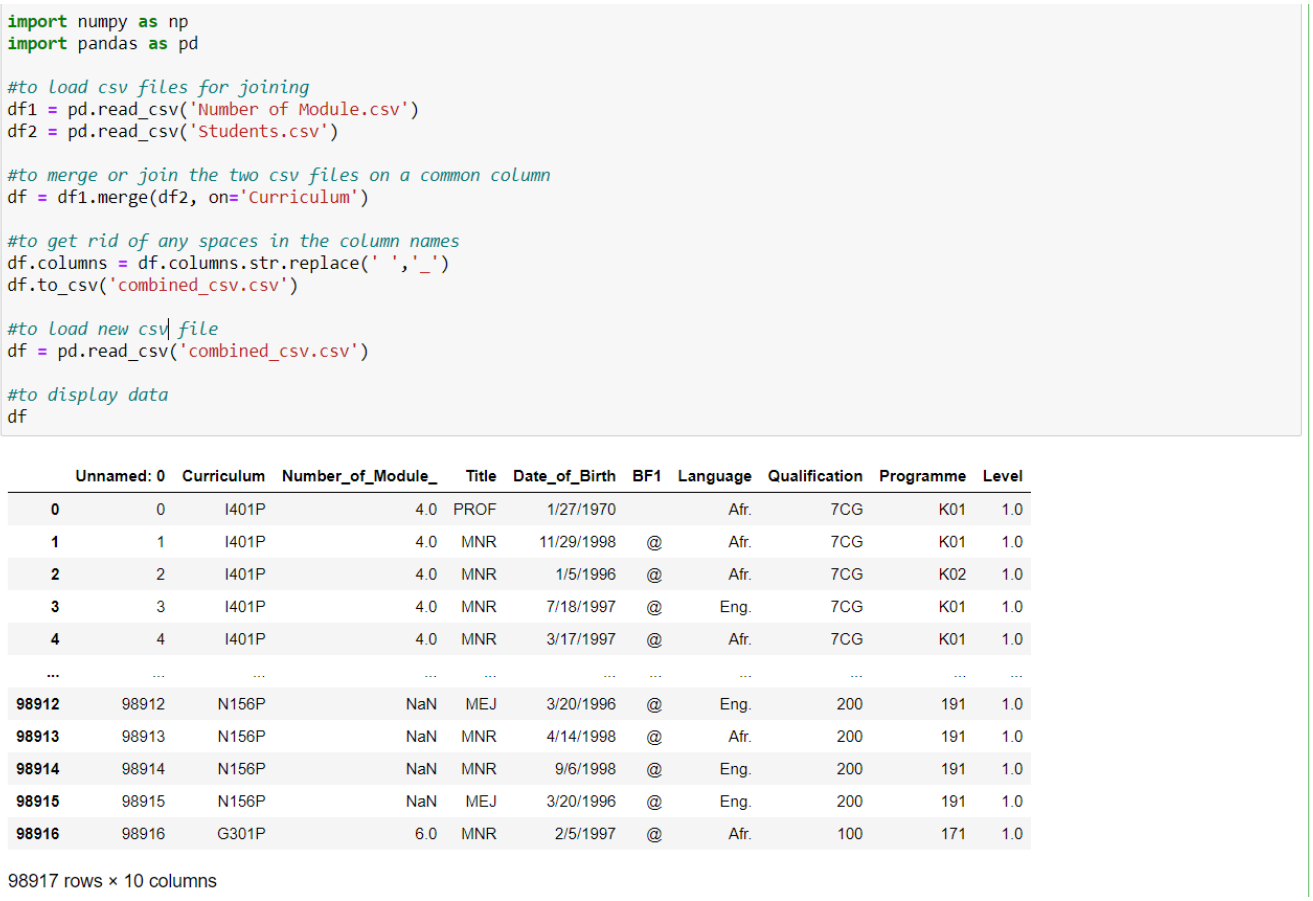
1. Find and replace the missing values (e.g.: NaN) with (#) from combined_csv. csv.
2. Extract and display all curriculum with 5 module from the combined_csv. csv.
3. Record a 5 minutes video where you explain how your program is working

Solutions to Question 3

Code to join the 2 csv files on a common column:

```
import numpy as np
import pandas as pd

#to load csv files for joining
df1 = pd.read_csv('Number of Module.csv')
df2 = pd.read_csv('Students.csv')

#to merge or join the two csv files on a common column
df = df1.merge(df2, on='Curriculum')

#to get rid of any spaces in the column names
df.columns = df.columns.str.replace(' ','_')
df.to_csv('combined_csv.csv')

#to load new csv file
df = pd.read_csv('combined_csv.csv')

#to display data
df
```

output:



```
import numpy as np
import pandas as pd

#to load csv files for joining
df1 = pd.read_csv('Number of Module.csv')
df2 = pd.read_csv('Students.csv')

#to merge or join the two csv files on a common column
df = df1.merge(df2, on='Curriculum')

#to get rid of any spaces in the column names
df.columns = df.columns.str.replace(' ','_')
df.to_csv('combined_csv.csv')

#to load new csv file
df = pd.read_csv('combined_csv.csv')

#to display data
df
```

	Unnamed: 0	Curriculum	Number_of_Module_	Title	Date_of_Birth	BF1	Language	Qualification	Programme	Level
0	0	I401P	4.0	PROF	1/27/1970		Afr.	7CG	K01	1.0
1	1	I401P	4.0	MNR	11/29/1998	@	Afr.	7CG	K01	1.0
2	2	I401P	4.0	MNR	1/5/1996	@	Afr.	7CG	K02	1.0
3	3	I401P	4.0	MNR	7/18/1997	@	Eng.	7CG	K01	1.0
4	4	I401P	4.0	MNR	3/17/1997	@	Afr.	7CG	K01	1.0
...	...	...	...	...	...	...	...	...	...	...
98912	98912	N156P	NaN	MEJ	3/20/1996	@	Eng.	200	191	1.0
98913	98913	N156P	NaN	MNR	4/14/1998	@	Afr.	200	191	1.0
98914	98914	N156P	NaN	MNR	9/6/1998	@	Eng.	200	191	1.0
98915	98915	N156P	NaN	MEJ	3/20/1996	@	Eng.	200	191	1.0
98916	98916	G301P	6.0	MNR	2/5/1997	@	Afr.	100	171	1.0

98917 rows × 10 columns

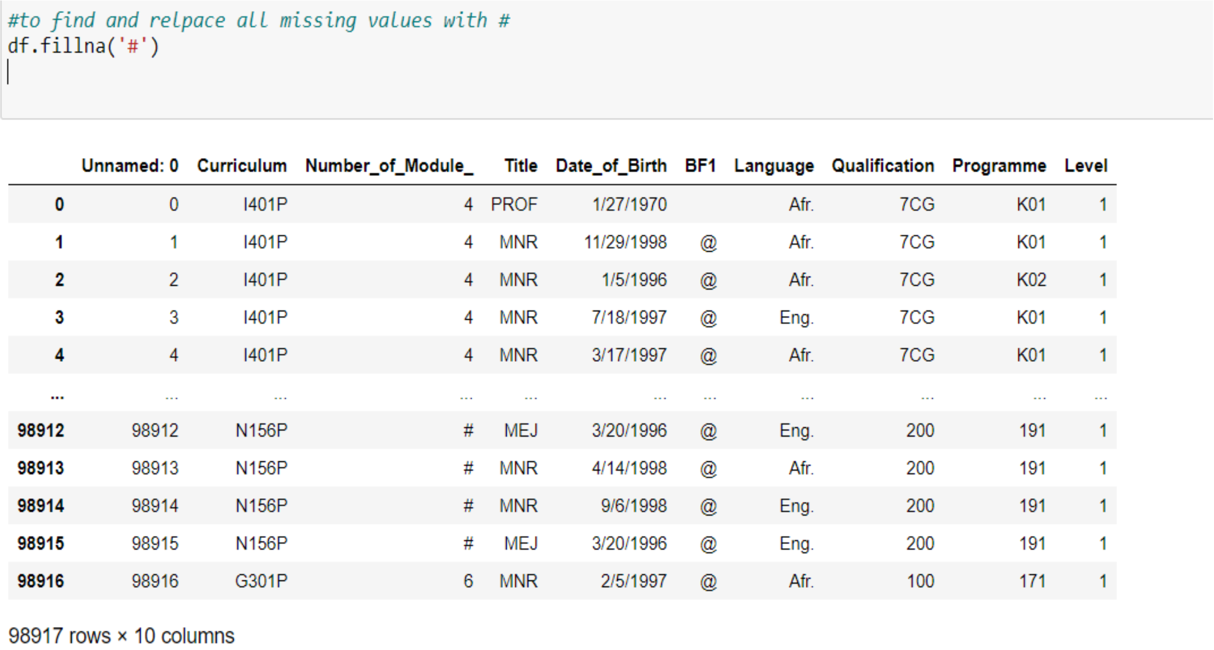
Question 1 And solution:

Find and replace the missing values (e.g.: NaN) with (#) from combined_csv. Csv

Code:

```
#to find and relpace all missing values with #
df.fillna('#')
```

output:



```
#to find and relpace all missing values with #
df.fillna('#')
```

	Unnamed: 0	Curriculum	Number_of_Module_	Title	Date_of_Birth	BF1	Language	Qualification	Programme	Level
0	0	I401P	4	PROF	1/27/1970		Afr.	7CG	K01	1
1	1	I401P	4	MNR	11/29/1998	@	Afr.	7CG	K01	1
2	2	I401P	4	MNR	1/5/1996	@	Afr.	7CG	K02	1
3	3	I401P	4	MNR	7/18/1997	@	Eng.	7CG	K01	1
4	4	I401P	4	MNR	3/17/1997	@	Afr.	7CG	K01	1
...	...	...	...	...	...	...	...	...	...	...
98912	98912	N156P	#	MEJ	3/20/1996	@	Eng.	200	191	1
98913	98913	N156P	#	MNR	4/14/1998	@	Afr.	200	191	1
98914	98914	N156P	#	MNR	9/6/1998	@	Eng.	200	191	1
98915	98915	N156P	#	MEJ	3/20/1996	@	Eng.	200	191	1
98916	98916	G301P	6	MNR	2/5/1997	@	Afr.	100	171	1

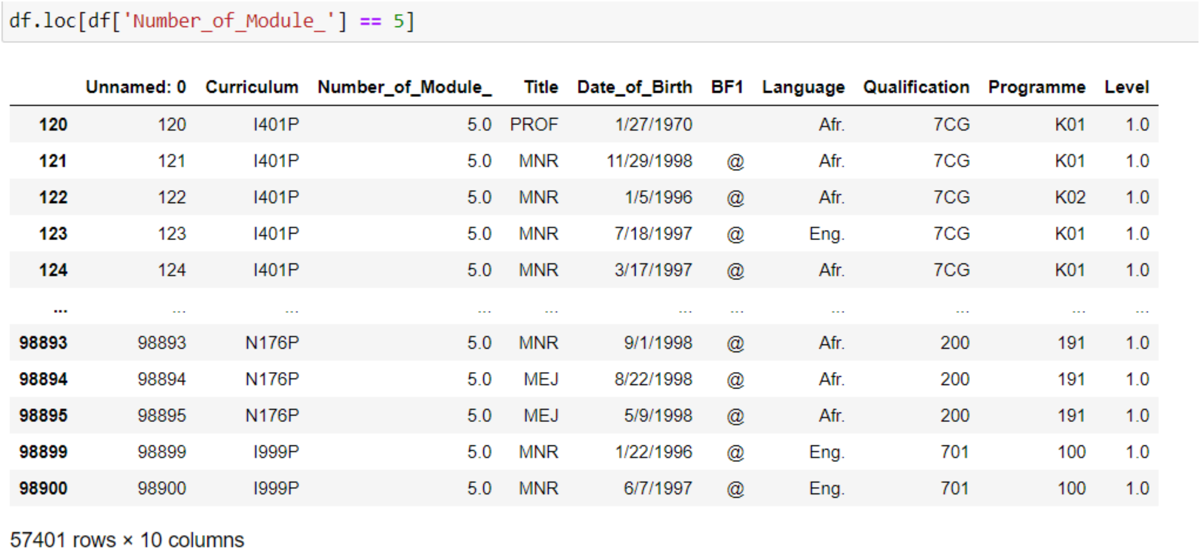
98917 rows × 10 columns

Question 2 and solution:
Extract and display all curriculum with 5 module from the combined_csv. csv.

Code:

```
df.loc[df['Number_of_Module_'] == 5]
```

Output:



```
df.loc[df['Number_of_Module_'] == 5]
```

	Unnamed: 0	Curriculum	Number_of_Module_	Title	Date_of_Birth	BF1	Language	Qualification	Programme	Level
120	120	I401P	5.0	PROF	1/27/1970		Afr.	7CG	K01	1.0
121	121	I401P	5.0	MNR	11/29/1998	@	Afr.	7CG	K01	1.0
122	122	I401P	5.0	MNR	1/5/1996	@	Afr.	7CG	K02	1.0
123	123	I401P	5.0	MNR	7/18/1997	@	Eng.	7CG	K01	1.0
124	124	I401P	5.0	MNR	3/17/1997	@	Afr.	7CG	K01	1.0
...	...	...	...	...	...	...	...	...	...	...
98893	98893	N176P	5.0	MNR	9/1/1998	@	Afr.	200	191	1.0
98894	98894	N176P	5.0	MEJ	8/22/1998	@	Afr.	200	191	1.0
98895	98895	N176P	5.0	MEJ	5/9/1998	@	Afr.	200	191	1.0
98899	98899	I999P	5.0	MNR	1/22/1996	@	Eng.	701	100	1.0
98900	98900	I999P	5.0	MNR	6/7/1997	@	Eng.	701	100	1.0

57401 rows × 10 columns

Question 3 and solution:
Record a 5 minutes video where you explain how your program is working

Link to Recording:
https://1drv.ms/v/s!AuRvf69NQWuRj19fff_EU6qODE7L?e=dgNzbU

END OF SECTION 3

Conclusions

In closing I hope this report succeeds in highlighting the a few of the more important skills needed as data scientist and how they are performed. When we bring all 3 sections together, we get a full view of how all these different skills and functions covered in the report work as one mechanism with each part heavily relying on the other in order to supplement a business need. Furthermore I hope that this report has proven to the read my capability to efficiently execute the role of a Data scientist.

Thank you.