Damian Vather

Explore Data Science Academy

Workbook Evidence

ITEM	RESULT
EDSA Orientation Quiz	75%
Analyse Skills Test	100%
Explain Skills Test	86%
Gather Skills Test	90%
Deploy Skills Test	60%
Basic Python MCQ	73%
Distributions and Sample Statistics MCQ	70%
Git, Shell, and Bash MCQ	100%
Intro to Probability MCQ	90%
Basic SQL Queries	100%
Set Theory	70%
Joins	100%
Modifying Data in SQL	90%
Writing Optimised SQL Queries	90%
Gather Predict MCQ	100%
Power BI Practical	100%
Explain Predict MCQ	96%
Power BI and Visualisation Theory	80%
Hypothesis Testing 2020	100%
Simple Linear Regression 2020	100%

Model Selection and Metrics 2020	100%
Regression MCQ exam 2020	85%
Classification theory test 2020	70%
NLP theory test 2020	90%
Dimensionality Reduction MCQ	76%
Clustering MCQ	70%
Unsupervised Exam	100%
Workplace Conduct MCQ	100%
Practitioners Meeting MCQ	100%
TOTAL	**88.31%**

Student: Damian Vather

Programme: Data Science

Test: EDSA Orientation Quiz : 26 Jan 2020

Mark: 75.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Below Average Test performance
You got less than 80% for the Test:

Fundamentals : Data Science 101 : EDSA Orientation Quiz

Q1:
I have read through, understood, and agree to abide by both the EDSA Honour Code and the EDSA Terms and Conditions. [1]

√ A: True

B: False

Q2: I have joined the EDSA Slack workspace. [1]

√ A: True

B: False

Q3: I have logged into my Explore Udemy account. [1]

A: False

√ B: True

Q4: All compulsory tutorials can be found under which tab on Athena? [1]

A: Test

B: Train

C: Pre-processing

X D: Problem Statement

Q5: The outline for each sprint can be found under which tab on Athena? [1]

A: Pre-processing

B: Problem Statement

C: Test

X D: Train

Q6: Under which tab should I submit my project deliverables on Athena? [1]

√ A: Predict

B: Pre-processing

C: Problem Statement

D: Test

Q7: Which of the following is not part of the Explore Data Science Process? [1]

A: Gather

B: Analyse

C: Explain

D: Deploy

√ E: Define

Q8: If I come across a word/concept that I don't understand, what should I not do? [1]

A: Ask my friends.

B: Ask a staff member.

C: Google it.

√ D: Give up.

Student: Damian Vather

Programme: Data Science

Test: Analyse Skills Test :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: Who created the Python programming language? (Google is your friend) [1]

A: Steve Jobs

B: Bill Gates

C: Elon Musk

√ D: Guido Van Rossum

Q2: What is the latest stable release of Python? [1]

A: 4.2

B: 2.4

C: 3.6

√ D: 3.8

Q3: What type of file extension does a Python script have? [1]

A: .ipynb

B: .pyth

√ C: .py

D: .python

Q4: What Python code would we need to add two variables, a and b, together? [1]

A:

a ++ b

B:

a add() b

C:

add(a, b)

D:

√ a + b

Q5: What Python code would we need to calculate a^2? [1]

A:

pow(a, 2)

B:

√ a**2

C:

a^2

D:

square(a)

Q6: What values can a variable of type “bool” take on? [1]

√ A: True or False

B: Any whole number

C: Text

D: Any decimal number

Q7: Which of the following is not a Python data type? [1]

A:

float

B:

int

C:

√ long

D:

str

Q8: To find out what data type variable a is, what Python function do we use? [1]

A:

who_is(a)

B:

var(a)

C:

√ type(a)

D:

what_is(a)

Q9: In Python, a list can contain multiple data types. [1]

A: False

√ B: True

Q10: If we have a list called “my_list”, how do we retrieve the 3rd element of that list? [1]

A:

my_list(2)

B:

my_list(3)

C:

√ my_list[2]

D:

my_list[3]

Student: Damian Vather

Programme: Data Science

Test: Explain Skills Test :

Mark: 85.71 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: What is Tufte's first rule of data visualization? [1]

A: Utilise colour

√ B: Show your data

C: Use graphics

D: Use labels

Q2: What is Tufte's 10th rule of data visualization? [1]

√ A: Understand narrative

B: Use labels

C: Utilise data ink

D: Use colour

Q3: What software will we be using to create dashboards this year at EDSA? [1]

A: Tableau

√ B: Microsoft Power BI

C: Matplotlib

D: Microsoft Excel

Q4: How do we import Matplotlib into Python? [1]

A:

import pyplot

B:

√ import matplotlib.pyplot as plt

C:

get matplotlib

D:

import matplotlib

Q5: How do we create a line plot in matplotlib? [1]

A:

plt.show()

B:

plt.line()

C:

plt.make()

D:

√ plt.plot()

Q6: How do we create a scatter plot in matplotlib? [1]

A:

plt.plot()

B:

plt.scatter_plot()

C:

√ plt.scatter()

D:

plt.show()

Q7: How do we create a histogram in matplotlib? [1]

A:

plt.plot()

B:

plt.histogram()

C:

plt.show()

D:

√ plt.hist()

Q8:
If we wanted to show the relationship between the two variables, Age and Height, which plot would be best? [1]

A: Line plot

√ B: Scatter plot

C: Histogram

D: Relationship plot

Q9: If we wanted to show the change in global population levels over time, which plot would be best? [1]

A: Scatter plot

X B: Histogram

C: Time plot

D: Line plot

Q10:
If we wanted to show the distribution of EDSA students across various education levels (high school, undergraduate, postgraduate), which plot would be best? [1]

A: Category plot

B: Histogram

C: Scatter plot

X D: Line plot

Q11: Which Google App is useful for storing and sharing files? [1]

A: Hangouts

B: Docs

C: Sheets

√ D: Drive

Q12: Which Google App is useful for writing reports? [1]

A: Word

B: Slides

√ C: Docs

D: Sheets

Q13: Which Google App is useful for creating slideshows? [1]

√ A: Slides

B: Docs

C: PowerPoint

D: Sheets

Q14: Which Google App is useful for creating spreadsheets? [1]

A: Docs

√ B: Sheets

C: Excel

D: Slides

Student: Damian Vather

Programme: Data Science

Test: Gather Skills Test :

Mark: 90.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: What does SQL stand for? [1]

√ A: Structured Query Language

B: Standard Query Language

C: Simplified Query Labels

D: Syntax Question Language

Q2: What version of SQL will we be using this year at EDSA? [1]

X A: PostgreSQL

B: MySQL

C: Microsoft SQL

D: Oracle

Q3: In a relational database, attributes would typically be recorded in: [1]

A: Tables

B: Rows

C: Chairs

√ D: Columns

Q4:
If we were to create a relational database to store personal records, it would typically be best practice to store each record as a: [1]

A: Table

B: Database

√ C: Row

D: Column

Q5: To select all columns from a table named “albums”, which of these queries would be correct? [1]

A:

GRAB * FROM albums;

B:

SELECT all FROM albums;

C:

FROM albums SELECT *;

D:

√ SELECT * from albums;

Q6: To select the “author” column from a table named “books”, which of these queries would be correct? [1]

A:

GET author FROM books;

B:

SELECT * FROM books;

C:

SELECT author FROM *;

D:

√ SELECT author FROM books;

Q7: Which keyword is used to return the total number of rows in a table? [1]

A:

ROWS

B:

LENGTH

C:

COUNT

√

D:

SUM

Q8: Which keyword is used to add a filter to a query? [1]

A:

WHEN

B:

WHERE

√

C:

IF

D:

ONLY

Q9: Which keyword(s) is/are used to sort the results of a query? [1]

A:

ORDER BY

√

B:

SORT

C:

ORDER

D:

SORT BY

Q10: If we wanted to know the number of films released per year, which query would we execute? [1]

A:

SELECT year FROM films GROUP BY year

B:

SELECT COUNT(*) FROM films GROUP BY year

C:

SELECT year, COUNT(*) FROM films

D:

SELECT year, COUNT(*) FROM films GROUP BY year

√

Student: Damian Vather

Programme: Data Science

Test: Deploy Skills Test :

Mark: 60.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Below Average Test performance
You got less than 80% for the Test:

Fundamentals : Data Science 101 : Deploy Skills Test

Q1: Which of the following is not a Project Management methodology? [1]

A: Waterfall

B: Agile

√ C: Kendo

D: Six Sigma

Q2: Which of the following is an AWS database service? [1]

A: S3

B: EC2

√ C: RDS

D: ELB

Q3: Software composed of smaller independent pieces are more difficult to deploy. [1]

A: False

X B: True

Q4: Which of the following is not commonly used as part of Agile methodology [1]

A: Kanban boards

X B: Rituals

C: MVPs

D: Sprint shuffles

Q5: Which tool will you be making use of in Deploy? [1]

A: AWS

X B: Git and Github

C: Trello

D: All of these tools

Q6: AWS is a distributed version control system (DVCS) [1]

A: False

X B: True

Q7:
Project Management and Version Control are key skills to produce and eventually deploy a product that will live in the real-world [1]

√ A: True

B: False

Q8: Which cloud service does EDSA mainly make use of? [1]

A: Google Cloud

B: Microsoft Azure

√ C: AWS

D: Kamatera

Q9: Which of the following is an important Deploy concept? [1]

√ A: Version control

B: Mathematics

C: Data storage

D: Business intelligence

Q10: Which of the following is not a project board best practice? [1]

A: Track single task using checklists

√ B: Always @mention the entire team so that everyone is on the same page

C: Use labels to signify the state of a card

D: Only assign someone to a card where they need to action something

Student: Damian Vather

Programme: Data Science

Test: Basic Python MCQ :

Mark: 73.33 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Below Average Test performance
You got less than 80% for the Test:

Fundamentals : Analyse : Basic Python MCQ

Q1: Which of the following is the best example of an `int` in Python? [1]

A:

"1"

B:

2

√

C:

05

D:

2.0

Q2: What does the following code output: str(9+1)? [1]

A:

X 10

B:

'9+1'

C:

9+1

D:

'10'

Q3: What does the following code output: 1 + 2 + 3.0? [1]

A:

√ 6.0

B:

6

C:

'3 + 3.0'

D:

'1 + 2 + 3.0'

Q4: In the Python list:

students = ["lindiwe", "tshepo", "rendani", "aaron", "michael", "nkhuna", "ngwato"]

Which of the following lines of code will print out "rendani"? [1]

A:

√ students[2]

B:

students(3)

C:

students(2)

D:

students[3]

Q5: Suppose we are given two Python lists: scores_1 = [2, 4, 3, 8]
scores_2 = [4, 4, 5, 3]

What will the output be if we run `scores_1 + scores_2`? [1]

A:

33

B:

[2, 4, 4, 4, 3, 5, 8, 3]

C:

√ [2, 4, 3, 8, 4, 4, 5, 3]

D:

[6, 8, 8, 11]

Q6:
Which of the following lines will output a new list sorted in alphabetical order from a list called students? [1]

A:

sort(students)

B:

sort[students]

C:

√ sorted(students)

D:

sort()

Q7: Which of the following lines will output the number of elements in a list called students? [1]

A:

length()

B:

√ len(students)

C:

len[students]

D:

length(students)

Q8: What is the FIRST string that the following code would print out in Python?

```
students = ["janneman", "jordan", "john", "jacob", "james"]

for i in range(students):
    if students[i][-1] != 'n' and students[i][-1] != 'b':
        print(students[i])
```

[1]

A: janneman

B: jordan

√ C: There will be an error

D: james

Q9: The difference between a list and tuple is that a tuple cannot be sliced, while a list can. [1]

√ A: False

B: True

Q10: A "while loop" is another name for a "for loop" [1]

A: True

√ B: False

Q11: Which one of the following statements about lists is false? [1]

A: The first item in a list starts at index 0

√ B: Lists can only consist of the same data type

C: Lists are ordered

D: Lists are mutable

Q12: Which code cannot be used to add the number 10 to the end of a list defined as a = [58, 45, 12]? [1]

A:

extend(10)

B:

extend([10])

X

C:

insert(3, 10)

D:

append(10)

Q13:

Which list comprehension statement can be used to replace the following code where a = [92, 56, 78, 41, 87]?

```
a_new = []
for i in a:
    if i < 70:
        a_new.append(i + 10)
    else:
        a_new.append(i + 5)
```

[1]

A:

a_new = [i + 10 if i < 70 else i + 5 for i in a]

B:

a_new = [if i < 70 then i +10 else i + 5 for i in a]

C:

a_new = [if i < 70 return i +10 else i + 5 for i in a]

D:

a_new = [for i in a return i + 10 if i < 70 else i + 5]

X

Q14: Which one of the following statements about sets is false? [1]

A: Items in a set are unique

B: Sets are ordered

C: Sets are mutable

X D: Every item in a set must be immutable

Q15: Which method allows us to access both key and value on a dictionary at the same time? [1]

A:

√ items()

B:

keys_values()

C:

keys()

D:

values()

Student: Damian Vather

Programme: Data Science

Test: Distributions and Sample Statistics MCQ :

Mark: 70.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Below Average Test performance
You got less than 80% for the Test:

Fundamentals : Analyse : Distributions and Sample Statistics MCQ

Q1: Which statement below best describes inferential statistics: [1]

A: Inferential statistics is the main analysis of machine learning

B: Inferential statistics is used to formulate statistical models which make conclusions about sample statistics

√ C: Inferential statistics is used to make predictions or comparisons about a larger group (a population) using information gathered about a small part of that population

D: Inferential statistics is a branch of sample statistics

Q2:
Which of these statements about data types and variables are correct:
i) Lap times are continuous data types
ii) Number of doctors your apple kept away today is a discrete variable
iii) Coming second in a relay race is a Nominal binary data type
iv) Genders are categorical data [1]

√ A: i), ii), iv)

B: i), iv)

C: i)

D: i), ii), iii), iv)

Q3: Which one of the following concerning sample and population statistics are correct: [1]

A: r represents sample regression and b represents population regression

B: r represents sample correlation and ? (rho) represents population correlation

C: Sample mean is denoted by the greek letter mu

X D: Square-rooting the population standard deviation will give you sample variation

Q4:
Which of the following mentioned standard Probability density functions is applicable to discrete Random Variables? [1]

A: Rayleigh

B: Gaussian

√ C: Poisson

D: Normal

Q5: Which statement about cumulative distributive functions (CDF's) is correct? [1]

A: The area under a conditional CDF is equal to 1

X B: CDFs for an exponential distribution and a normal distribution look the same

C: Denoted by F(p(x))

D: Give the probability that a random variable X takes on a value greater than or equal to some given value x

Q6:
You draw a random sample of size n = 64 from a population with mean μ = 50 and standard deviation σ = 16. From this, you compute the sample mean, $\bar{X}$. What is the standard deviation deviation of $\bar{X}$? [1]

A: 5.4

B: 1.2

C: 3

√ D: 2

Q7:
The number of belt points you miss follows a Poisson distribution. These belt points are missed independently at an average rate of 1.65 points per belt tier. What is the probability that you did not miss any belt points in 3 belt tiers (white,yellow orange). [1]

A: 1.060

B: 1.008

C: 0.0060

√ D: 0.0071

Q8: Fill in the missing inputs (where ___ is present) in the following code for plotting a normal distribution:

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
%matplotlib inline

mu = 0
variance = 2
sigma = np.___(variance) ,

x = np.linspace(mu-3\*variance,mu+3\*variance, 100)

plt.plot(x, ___.pdf___, mu, sigma))
```

[1]

A: log, norm, (x^3)

B: sqrt, poisson, x^3

√ C: sqrt, norm, x

D: sqrt, poisson, x*mu

Q9: Pick the incorrect answer concerning the stats package scipy.stats: [1]

A: You can calculate only the mean of your data

X B: The package has a function to use certain hypothesis testing techniques based on input data

C: When passing a vector through a function, it can output a CDF or PDF at various points

D: The package has a function which will output a string to tell you if your hypothesis is rejected or not

Q10:
The completion time of completing this test is normally distributed with mean of 65 minutes and a standard deviation of 23 minutes. What is the probability that the completion time will be between 40 and 100 minutes?
Hint : Use Normal distribution formulas. When using the z-table, you will need to use a z-table with positive z-scores and another for negative z-scores to find two probabilities. [1]

A: 75%

√ B: 80%

C: 65%

D: 70%

Student: Damian Vather

Programme: Data Science

Test: Git, Shell, and Bash MCQ :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: In a bash script, what does the command "echo" do? [1]

√ A: Outputs the input on a new line

B: Replaces the "ping" command used in Linux Systems

C: Executes a block of code

D: Allows the programmer to store commands in the shell

Q2: In the windows CMD shell, what will the "cd" command do? [1]

A: Cache directory refresh

B: Convert directory folder

√ C: Change directory folder

D: Change directory of the file

Q3: In the windows CMD shell, how would you access the root directory? [1]

A:

√ cd C:/

B:

cd

C:

cd -home

D:

cd..

Q4:
Which of the following statements are true? i) "ls" will list all the files and folders inside a directory in bash script
ii) DIR will list all the files and folders inside a directory in Windows CMD
iii) rm file1 file2 will move and rename file1 to a new directory as file2
iv) shell and bash scripts are stored with .sh extensions [1]

A: i), ii), iii), iv)

B: i), iii)

√ C: i),ii), iv)

D: ii), iv)

Q5:
Which of the following statements concerning GIT are true: i) GIT is a distributed version control system
ii) GIT can be used to push changes onto a remote repository on Github using a push command and to download a remote repository using a pull command
iii) In Github, you can create a subfolder by adding a "/" when creating/editing a repository
iv) GIT can be accessed through bash commands [1]

√ A: All of these

B: i), ii), iv)

C: i), iv)

D: i), ii)

Q6:
In bash script, if you run the following commands, what will happen (Assuming you are in the user directory)?

cd Documents
mkdir python_boss
cd python_boss
ls

[1]

A: Change the directory to Documents, make a new directory with a file called "python_boss", change the directory to "python_boss" and list all the items in that file

B: Change the directory to Documents, make a new file called "_boss", and list all the items in that file

C: Change the directory to /c/Documents, make a new directory called "python_boss", change the directory to "python_boss", and list all the documents in the folder

√ D: Change the directory to Documents, make a new directory called "python_boss", change the directory to "python_boss", and list all the documents in the folder

Q7: In relation to bash script, which one of these statements is incorrect: [1]

A: Linux uses bash scripting

B: sudo will run a command with administrative privileges

√ C: AWS runs on a Windows style CMD shell, but using bash commands

D: You can access an AWS instance using ssh

Q8: In bash scripting, you run the following commands inside the home directory. What is the outcome?

```
sudo mkdir test
cp file1 test
```

[1]

A: Creates a directory under administrator user rights called "test" and copies file1 into the test directory as a backup

B: Builds a new file directory called "test" as a superuser to allow for file1 to be copied to test as a backup

C: Creates a file in the Documents directory called "test" and changes the file name of "test" to "file1"

√ D: Creates a directory under administrator user rights called "test", and copies file1 to that directory

Q9: After executing the following commands in bash, what will happen:

```
cd ..
pwd
```

[1]

A: Create a folder called pwd

B: It will give an error

√ C: Change to parent directory and display the path of the current directory

D: You'll be prompted to change your password.

Q10: Say you wanted to create a new git repository locally. Which command would you use? [1]

A:

git repository -n

B:

git init -locally

C:

√ git init

D:

git repository

Q11: Which of the following commands will execute a python script in bash? [1]

A:

√ python path/py_script.py

B:

python -path/py_script.py

C:

python -path/py_script.ipynb

D:

py path/py_scrpit.py

Q12: When you first set-up your git account, which command do you use to create a user name? [1]

A:

√ git config --global user.name "yourusername"

B:

git configure user.name "yourusername"

C:

git config user.name "yourusername"

D:

git configure --global user.name "yourusername"

Q13: How would you stage a README.md file to your git project? [1]

A:

git commit README.md

B:

git clone README.md

C:

√ git add README.md

D:

git push README.md

Q14: Which best describes a copy of a git repository? [1]

A:

git branch

B:

git copy

C:

git master

D:

√ git fork

Q15: What is the difference between git fetch and git pull? [1]

A: There is no difference.

B: Git fetch retrieves a list of changes from a git repo. Git pull will submit your local changes to a repo.

C: Git fetch clones a git repo to your local directory. Git pull will merge changed files with your local copy.

√ D: Git fetch will copy the metadata of changed files from a repo, but will not alter your copy. Git pull will do that and make changes to your local copy.

Student: Damian Vather

Programme: Data Science

Test: Intro to Probability MCQ :

Mark: 90.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1:
Which one of the following statements are correct? i) An independent event is an event which has no related causality on another event
ii) The chance of rolling a 2 on a dice is 2/6
iii) The sample space for tossing a coin can be denoted as S = {H;T} [1]

A: ii), iii)

X B: All are correct

C: i)

D: i), iii)

Q2: A dependent event can be described as which of the following scenarios: [1]

√ A: Having a bag of pool balls and finding the probability of picking out 2 striped balls without replacement

B: Picking a red ace card from a deck of cards

C: Achieving a top 10 position at Explore regardless of your performance

D: Rolling a dice

Q3: Probability is: [1]

A: Not a changing variable when the possibility of events occuring increase

B: Mutually exclusive

C: The sum of all the probabilities of possible events occurring which are all greater than 1

√ D: The likelihood that an event will occur

Q4:
You toss a fair coin 6 times. How would you calculate the likelihood of getting ('H', 'H', 'H', 'H', 'H', 'H'), or six heads? [1]

√ A: Multiply the probability of getting heads by itself 6 times.

B: Add the probability of getting heads 6 times.

C: Take a sample of 1000 coin tosses and measure the frequency of 6 heads

D: Use a gaussian distribution over a single coin toss.

Q5: The probability of event A happening given B has already happened is: [1]

√ A: A conditional probability

B: Always equal to 1

C: P(A) - P(B)

D: An equally likely event

Q6:
You want to simulate ten rolls of a die. You proceed to programme this into python. Fill in the missing blanks (as __) in the following piece of code to complete the die rolling program:

```
import numpy as np
outcomes = np.random._____(1,__, ___=10)
print(outcomes)
```

[1]

A:

random, 6, batch

B:

randint, 6, batch_size

C:

random, 7, len

D:

√ randint, 7, size

Q7: You flip a coin twice. What is the probability of getting a tails on both tosses? [1]

A: 1/8

√ B: 0.25

C: 1/2

D: 2/3

Q8:
You are writing a python exam. You make a bet with your Supervisor that you will pass the exam. You also bet that out of your entire group (4 people total, including you), that you will be the only one to pass. What is the probability that you will win this bet, assuming that everyone in the group has a 75% chance of passing? [1]

A: 1/16 chance

√ B: 3/256 chance

C: 1/2 chance

D: No chance at all

Q9:
Given the scenario in the previous question, what is the probability that any 3 members of your group will pass the exam? [1]

A: 27/256

B: 1/4

√ C: 27/64

D: 1/3

Q10:
The Joburg campus did blood tests for health awareness week. There are 150 students. The results came back as follows:

- A types :25 students are A+, 28 are A-
- B Types: 14 students B+, 7 are B-
- AB types: 8 students are AB+, 0 are AB-
- O Types: 38 students are O+, 30 are O-

What is the probability of being an A type blood? [1]

√ A: 0.353

B: 0.406

C: 0.167

D: 0.187

Student: Damian Vather

Programme: Data Science

Test: Basic SQL Queries :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1:
The questions should be answered based on the TMDb database. Students should be very familiar with the database by the time of answering this test. All of the questions and answers should be based only on the Movies table. To attempt this test it would be beneficial to know the following functions and statements: SELECT, FROM, WHERE, LIKE, COUNT, OR, AND, BETWEEN. What is the SQL code to see the whole movies table: [1]

A:

√ SELECT * FROM Movies

B:

SELECT TABLE FROM Movies

C:

SELECT ALL FROM Movies

D:

SELECT % FROM Movies

Q2: What was the budget for the movie “Inception” [1]

A: $115 000 000

√ B: $160 000 000

C: $344 000 000

D: $224 000 000

Q3: What is the runtime of the movie Titanic? [1]

A: 158

B: 122

√ C: 194

D: 146

Q4:
How many movies do not have English as their original language? (Hint: “en” is the abbreviation for English) [1]

√ A: 298

B: 387

C: 315

D: 492

Q5: How many movies are there that have a popularity score of more than 250? [1]

√ A: 7

B: 5

C: 11

D: 9

Q6: How many movies are there where the title is not the same as the original title? [1]

A: 74

B: 24

√ C: 258

D: 187

Q7:
How many movies are there that managed to get a popularity rating of more than 100 with a budget of less than $10 000 000? [1]

√ A: 5

B: 8

C: 11

D: 15

Q8:
How many movies are there that has the word ‘love’ anywhere in the title? (Hint: The L in the word love can be upper or lower case and can be included in words such as ‘lovers’.) [1]

A: 58

B: 49

C: 67

√ D: 71

Q9: How many movies were released between the dates 1 August 2012 and 31 July 2013? [1]

√ A: 227

B: 3

C: 208

D: 295

Q10:
You have had a long day and want to sit back and enjoy a movie. Unfortunately, today you are only in the mood for a very specific type of movie. It definitely needs to be in English. It should also be new, something after 1 Jan 2010, but not too new as you might have seen it recently, so it must have been released before 1 Jan 2016. It should also be a romantic movie, so make sure it has the word love somewhere in the title. You also want it to be a big blockbuster, so the budget of the move must be more than $10 000 000. What is the movie with the highest popularity that meets all of your requirements? [1]

A: Love & Other Drugs

B: From Paris with Love

C: Eat Pray Love

√ D: Crazy, Stupid, Love

Student: Damian Vather

Programme: Data Science

Test: Set Theory :

Mark: 70.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Below Average Test performance
You got less than 80% for the Test:

Fundamentals : Gather : Set Theory

We recommend that you do the following Train/Pre-processing:

Fundamentals : Gather : Normalising Tables
Fundamentals : Gather : Querying a Database
Fundamentals : Gather : SQL & Database Design A-Z™: Learn MS SQL Server + PostgreSQL

Q1: The union of A and B is defined as all items that belong to either A or B, or both A and B. [1]

X A: False

B: True

Q2: How many students did not write any of their final exams? [1]

A: 95

B: 70

C: 25

√ D: 5

Q3:
How many students did not have to write exams on two separate occassions (i.e. only wrote finals or only wrote supps)? [1]

√ A: 30

B: 70

C: 95

D: 5

Q4: How many students had to re-write their maths exam? [1]

A: 5

X B: 30

C: 80

D: 20

Q5: How many students had to re-write their science exam? [1]

X A: 30

B: 20

C: 24

D: 23

Q6:
What was the average mark for students who wrote the supplementary accounting exam after missing the first? [1]

A: 82

√ B: 76

C: 79

D: 73

Q7:
What was the average mark for students who wrote the supplementary accounting exam after failing the first? [1]

A: 76

B: 79

√ C: 73

D: 82

Q8: What was the average mark for all students who wrote the final computer science exam? [1]

A: 76

B: 73

√ C: 67

D: 52

Q9: What was the average mark for all students who wrote the supplementary computer science exam? [1]

A: 76

√ B: 71

C: 73

D: 67

Q10:
Assuming all subjects are weighted equally, what was the average total mark for students who didn't write any supplementary exams? [1]

A: 66

B: 73

C: 76

√ D: 74

Student: Damian Vather

Programme: Data Science

Test: Joins :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: What is the primary key for the table “movies”? [1]

A: film_id

B: title

C: movie_key

√ D: movie_id

Q2: How many foreign keys does the “languagemap” table have? [1]

A: 3

√ B: 2

C: 0

D: 1

Q3: How many movies in the database were produced by Pixar Animation Studios? [1]

A: 20

B: 18

C: 14

√ D: 16

Q4:
What is the most popular action movie that has some German in it? (Hint: The German word for German is Deutsch) [1]

A: The Bourne Identity

B: Quantum of Solace

C: Mission: Impossible - Rogue Nation

√ D: Captain America: Civil War

Q5:
In how many movies did Tom Cruise portray the character Ethan Hunt? (Hint: Characters are listed in the Casts table.) [1]

A: 4

√ B: 5

C: 6

D: 3

Q6: How many times was the actress Cate Blanchett nominated for an Oscar? [1]

A: 2

√ B: 7

C: 4

D: 5

Q7:
How many movies have managed to win Best Picture at the Oscars even though they had a budget of less than $10 000 000? (Hint: The winner is given by a 1 in the “winner” field.) [1]

A: 12

B: 16

√ C: 15

D: 18

Q8: How many movies contain at least one of the languages, Afrikaans or Zulu? [1]

A: 10

B: 12

√ C: 8

D: 15

Q9: In which countries was the movie “Taken” produced? [1]

A: United States of America, Canada, Spain

√ B: United States of America, United Kingdom, France

C: Canada, Spain, France

D: United States of America, United Kingdom, Spain

Q10: How many movies are in the database that are both a Romance and a Comedy? [1]

√ A: 484

B: 373

C: 595

D: 262

Student: Damian Vather

Programme: Data Science

Test: Modifying Data in SQL :

Mark: 90.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1:
What query would you write to change the name of the language with the 'zh' iso code in the "Languages" table to 'Chinese'? [1]

A:

MODIFY Languages SET language_name = 'Chinese' WHERE iso_639_1 = 'zh'

B:

UPDATE Languages (language_name = 'Chinese') WHERE iso_639_1 = 'zh'

C:

√ UPDATE Languages SET language_name = 'Chinese' WHERE iso_639_1 = 'zh'

D:

ALTER Languages SET language_name = 'Chinese' WHERE iso_639_1 = 'zh'

Q2: What query would you write to insert a new genre called 'Sport' with an id of 10? [1]

A:

INSERT INTO Genres (genre_id = 10, genre_name = 'Sport')

B:

INSERT Genres (genre_id, genre_name) Values (10, 'Sport')

C:

INSERT (genre_id, genre_name) INTO Genres SET VALUE (10, 'Sport')

D:

√ INSERT INTO Genres (genre_id, genre_name) Values (10, 'Sport')

Q3:
You have just watched The Flintstones movie and did not find it very funny. What code would delete the entry that links The Flintstones to the Comedy genre? [1]

A:

√ DELETE FROM genremap WHERE genre_id = 35 and movie_id = 888

B:

REMOVE ENTRY FROM genremap WHERE genre_id = 35 and movie_id = 888

C:

DELETE FROM genremap (genre_id = 35, movie_id = 888)

D:

DELETE ENTRY FROM genremap WHERE genre_id = 35 and movie_id = 888

Q4:
What query would you write to add a column to the language table that could be used for the English names of the different languages? [1]

A:

ALTER TABLE languages ADD language_english_name

B:

√ ALTER TABLE languages ADD language_english_name varchar(50)

C:

UPDATE TABLE languages ADD language_english_name

D:

UPDATE TABLE languages APPEND language_english_name varchar(50)

Q5:
If the foreign keys in the “GenreMap” table are set to ON DELETE CASCADE, which of the following is true? [1]

A: Deleting an entry in the “GenreMap” table will lead to all entries with a matching “genre_id” in the “GenreMap” table to also be deleted.

B: Deleting an entry in the “Genre” table will lead to all entries with a matching “genre_id” in the “GenreMap” table to also be deleted.

X C: Deleting an entry in the “GenreMap” table will lead to all entries with a matching “genre_id” in the “Genre” table to also be deleted.

D: Deleting an entry in “GenreMap” will lead to no other changes in the data.

Q6:
What query would you write to create a new table called “Favourites” that only has one column for the movie_id? [1]

A:

CREATE TABLE Favourites with COLUMN(movie_id int)

B:

NEW TABLE Favourites(movie_id int)

C:

NEW TABLE Favourites with COLUMN(movie_id int)

D:

√ CREATE TABLE Favourites (movie_id int)

Q7:
If you want to make sure that you do not put duplicates in your new “Favourites” table, how would you change the table so that the movie_id column has to be unique? [1]

A:

√ ALTER TABLE Favourites ADD UNIQUE(movie_id)

B:

UPDATE TABLE FAVOURITES (movie_id int unique)

C:

ALTER TABLE FAVOURITES (movie_id int unique)

D:

UPDATE TABLE FAVOURITES ADD UNIQUE(movie_id)

Q8:
Let us say that you want your new “Favourites” table to hold the names of your favourite movies and not the movie_id’s. What query would you write to change the data type of the movie_id column from integer to text? [1]

A:

ALTER TABLE Favourites MODIFY COLUMN(movie_id varchar(500))

B:

√ ALTER TABLE Favourites ALTER COLUMN movie_id varchar(500)

C:

UPDATE TABLE Favourites CHANGE movie_id varchar(500)

D:

UPDATE TABLE Favourites MODIFY movie_id int TO varchar(500)

Q9: What would the following query do? INSERT INTO Favourites
SELECT title FROM Movies
WHERE popularity > 100; [1]

A:

Displays the names of all movies with a popularity of more than 100 but does not change anything in the tables.

B:

Inserts the names of the 100 most popular movies into the Favourites table.

C:

√ Inserts the names of all movies with a popularity of more than 100 into the Favourites table.

D:

Inserts the movie_id of all movies with a popularity of more than 100 into the Favourites table.

Q10: What query would you write to delete the table “Favourites” that we created in the previous question? [1]

A:

√ DROP TABLE Favourites

B:

DELETE TABLE Favourites

C:

PURGE TABLE Favourites

D:

REMOVE TABLE Favourites

Student: Damian Vather

Programme: Data Science

Test: Writing Optimised SQL Queries :

Mark: 90.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1:
The following 10 questions will be based on the USA domestic flights database for 2008. Download and restore the full Flights database, available in the "Optimising SQL Queries". Make sure you are NOT using the reduced database available in the Pyodbc trains.
How many different carriers are there in total in the database? [1]

A: 1 252

B: 37

√ C: 1 491

D: 20

Q2: How long was the longest delay before departure? [1]

A: 999 min

√ B: 2467 min

C: 2457 min

D: 588 min

Q3: How many flights departed on the 23rd of September? [1]

A: 540 908

√ B: 18 056

C: 35 124

D: 75 887

Q4: What is the distance between Midway Airport (MDW) and Houston Airport (HOU)? [1]

√ A: 937

B: 611

C: 972

D: 1121

Q5: Which month had the highest number of cancelled flights? [1]

A: December

B: September

C: August

√ D: February

Q6: How many airports have the word "International" in their name? [1]

A: 2

B: 8

C: 110

√ D: 124

Q7: What is the most produced model for the manufacturer "BOEING"? [1]

A: 757-222

B: 737-3H4

√ C: 737-7H4

D: 717-200

Q8: What manufacturer had the highest average delay time (DepDelay + ArrDelay)? [1]

A: PAIR MIKE E

B: AERONCA

√ C: BOEING OF CANADA LTD

D: DEHAVILLAND

Q9: How many flights were scheduled to land at Los Angeles International Airport? [1]

A: 41 258

√ B: 215 685

C: 215 608

D: 39 422

Q10:
Which domestic carrier had the best on-time performance (OTP), where OTP is defined as the rate of on time flights with a 15min buffer on departure and arrival? [1]

A: Comair Inc.

B: Hawaiian Airlines Inc.

C: Aloha Airlines Inc.

X D: American Airlines Inc.

Student: Damian Vather

Programme: Data Science

Test: Gather Predict MCQ :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: How many locations are there in the database? [1]

A: 26

√ B: 34

C: 28

D: 52

Q2: Which location has the most amount of stations? [1]

A: Pretoria

B: Rustenburg

C: Emalahleni

√ D: Middelburg

Q3: Which coastal province has the most number power plants? [1]

A: Kwa-zulu Natal

B: Mpumalanga

√ C: Western Cape

D: Eastern Cape

Q4: What is the second most common type of power station? [1]

√ A: Hydroelectric

B: Coal

C: Gas Turbine

D: Wind Power

Q5: Which of the following locations has the most number of hydroelectric stations? [1]

A: Norvalspont

√ B: Umtata River

C: Petrusville

D: Ncora River

Q6: According to the data provided, how many power plants are/were operated by Eskom? [1]

A: 25

√ B: 32

C: 28

D: 39

Q7: Which of these power stations have been decommissioned? [1]

A: Medupi, Limpopo

√ B: Klipheuwal, Western Cape

C: Kusile, Mpumalanga

D: Kusile, Limpopo

Q8: Which power station has the largest amount of power available for distribution (nominal capacity)? [1]

A: Lethabo

B: Matimba

C: Kendal

√ D: Majuba

Q9:
How would you select the number of houses electrified in the year 2008 in Mpumalanga in SQL? Select the SQL code below (assuming you did not change the name of the tables): [1]

A: select province_id_fk, year from provincial_household_electrified where id = 1 and year = 2008

B: select 2008 from provincial_household_electrified where province_id_fk = 1

√ C: select household_count, province_id_fk, year from provincial_household_electrified where province_id_fk = 1 and year = 2008

D: select province_id_fk, year from provincial_household_electrified where province_id_fk like 1 and year = '2008'

Q10:

10. Select the tables in the database (assuming you did not change the names) which best describe the supply and the demand of electricity:

i. national_electricity_generated_monthly
ii. population
iii. stations_meta
iv. stations_type_info [1]

√ A: Demand: i ; Supply: iii

B: Demand: i ; Supply: ii, iv

C: Demand: i,ii ; Supply: iv

D: Demand: iii ; Supply: i,ii, iv

Student: Damian Vather

Programme: Data Science

Test: Power BI Practical :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: Which batsman scored the most runs in 2013? [1]

A: CH Gayle

B: SR Tendulkar

√ C: MEK Hussey

D: JP Duminy

Q2: Which bowler had the highest extra runs in one game? [1]

A: SL Malinga

B: JA Morkel

C: SR Watson

√ D: DP Nannes

Q3:
In 2010, which batsman had the 2nd highest batting average and scored more than 100 runs against Kings XI Punjab? [1]

√ A: MJ Lumb

B: MK Tiwary

C: JH Kallis

D: R Dravid

Q4: In which year did DR Smith score the most runs (excluding extras)? [1]

A: 2010

B: 2017

C: 2013

√ D: 2014

Q5: Against which team did AM Nayar have the highest strike rate? [1]

A: Mumbai Indians

√ B: Chennai Super Kings

C: Delhi Daredevils

D: Rajasthan Royals

Q6: Who scored the 5th highest number of runs between 23 April 2015 and 21 May 2017? [1]

A: DA Warner

B: AB de Villiers

C: Mandeep Singh

√ D: G Gambhir

Q7: Against which bowler did V Kohli hit the most 6s in one game? [1]

A: R Ashwin

√ B: KC Cariappa

C: DJ Bravo

D: UT Yadav

Q8: Which batsman scored 4 centuries in 2016? (hint: a century is 100 runs in a single game) [1]

√ A: V Kohli

B: CH Gayle

C: S Raina

D: G Gambhir

Q9: Which team has the highest number of runs scored against them in a single game? [1]

A: Kings XI Punjab

B: Royal Challengers Bangalore

√ C: Pune Warriors

D: Mumbai Indians

Q10: What proportion of runs has CH Gayle scored in 6s over his IPL career for the RCB cricket team? [1]

A: 12.31%

B: 17.1%

√ C: 45.35%

D: 25.6%

Student: Damian Vather

Programme: Data Science

Test: Explain Predict MCQ :

Mark: 95.83 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: What is the total installed capacity in Middelburg? [1]

A: 8760 MW

B: 3525 MW

C: 1250 MW

√ D: 5235 MW

Q2: How many stations run by Bethlehem Hydro are operational? [1]

A: None are operational

B: 3

√ C: 2

D: 1

Q3: What is the total capacity available for distribution (nominal) by Eskom run power stations? [1]

√ A: 43762 MW

B: 43976 MW

C: 43652 MW

D: 43802 MW

Q4:
Which of the following fully operational stations has the largest difference between the total installed capacity and the total nominal capacity? [1]

A: Kusile

B: Duvha

C: Medupi

X D: Klipheuwel

Q5:
During the month of June, which year had the lowest electricity generation? Hint (electricity_generated_MW is the metric being evauluated) [1]

A: 2018

√ B: 2016

C: 2017

D: 2019

Q6:
Which of the following statements are true: i. Eastern Cape has only hydroelectric and gas/liquid fuel turbine stations
ii. The most common type of station in the Western Cape is Gas/liquid turbine
iii. Mpumulanga has only one hydroelectric station type
iv. The Northern Cape province has a total installed capacity of 610 MW and 4 hydroelectric stations [1]

A: i, iii, iv

B: ii, iv

C: i, iv

√ D: i, ii, iii

Q7: What is the total installed capacity (MW) for all provincial coastal Gas/ liquid fuel turbine stations? [1]

√ A: 2779 MW

B: 2761 MW

C: 2943 MW

D: 3449 MW

Q8: Which location with 2 operational coal stations has the highest combined installed capacity? [1]

A: Pretoria

√ B: Emalahleni

C: Lephalale

D: Bethal

Q9: Which of the following stations are operated by International Power? [1]

A: Kelvin

B: Kakamas

C: Newcastle

√ D: Avon

Q10:
Which of the following statements are correct: Hint (nominal_capacity_MW and installed_capacity are the metrics being considered) i) The Medupi power station has the largest installed capacity out of all the fully operational stations
ii) Lethabo station has the fourth highest capacity installed out of all the fully operational stations
iii) Komati and Avon Stations have nominal capacities of zero/null
iv) The Duvha power station has a better nominal capacity than the Matla power station [1]

A: i), ii), iii)

B: ii) only

C: ii), iv)

√ D: ii), iii)

Q11: In which province does eskom operate the most stations in [1]

A: Eastern Cape

B: Western Cape

√ C: Mpumalanga

D: Gauteng

Q12: In Gauteng, how many coal stations are still operational. [1]

√ A: 3

B: 2

C: 4

D: 1

Q13: You can create new visuals by dragging measures from the field pane straight onto the report page [1]

√ A: True

B: False

Q14: Which DAX formula could you use to replace one word or character in a text string with another? [1]

√ A: Substitute

B: Search

C: Find

D: Replace

Q15: Which one of the following data sources can Power BI NOT connect to? [1]

A: CSV file

B: JSON file

√ C: APK file

D: SQL Database

Q16: Which type of visual in Power BI would be best suited if we want to display data over time? [1]

A: Card

B: Funnel Chart

√ C: Line Chart

D: Bar Chart

Q17: Which of the following are not a type of filter available in Power BI Reports? [1]

A: Page level filter

B: Report level filter

√ C: Model level filter

D: Visual level filter

Q18: Report level filters only filters for the page it was created on? [1]

A: True

√ B: False

Q19: What does DAX in PowerBI stand for? [1]

A: Data Analytics Expressions

B: Direct Analytics Expressions

√ C: Data Analysis Expressions

D: Direct Analysis Expressions

Q20: What is the best way to represent the amount of electricity each province produces in PowerBI? [1]

A: Stacked bar graph

B: Line Graph

√ C: Bar Graph

D: Pie Chart

Q21: Which view would you use to see the relationships between your tables [1]

A: Report

√ B: Model

C: Data

D: Fields

Q22:
What is the best way to represent the contribution of the installed and nominal capacities of the stations per province to the total capacity of the grid? [1]

A: Doughnut Chart

B: Funnel Chart

√ C: Clustered Column Chart

D: Treemap

Q23: Which of the following would the COUNTA function count? [1]

A: Distinct values only

B: Non Numeric values only

C: Numerical values only

√ D: Non-blank cells (Numerical and Non-numerical)

Q24:
Which data connectivity mode should you use if you want the data to automatically refresh as you interact with your visuals [1]

√ A: DirectQuery

B: Import

Student: Damian Vather

Programme: Data Science

Test: Power BI and Visualisation Theory :

Mark: 80.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: Which of the following data sources can Power BI NOT connect to? [1]

A: CSV File

B: SQL Database

C: Excel File

√ D: Docx file

Q2: Which of these filtering methods can be used to filter data in Power BI?
1. Drillthrough filter
2. Dimensionality reduction filter
3. Report-level filter
4. Page-level filter [1]

A: Only 1

B: 1 , 2 and 3

√ C: 1, 3 and 4

D: 2 and 1

Q3: What is the best graph for displaying categorical data? [1]

A: Pie chart

B: Line graph

√ C: Bar graph

D: Scatter chart

Q4: Which of the following is not a principle of Edward Tufte? [1]

X A: Data-Ink

B: Data Intensity

C: Chartjunk

D: Data Density

Q5:
What is the result of the following DAX expression? <pre>Logical Test = IF(ISLOGICAL(“true”), "Is Boolean type or Logical", "Is different type")</pre> [1]

A: Logical Test is undefined

B: Is Boolean type or Logical

C: “true”

√ D: Is different type

Q6:
Suppose you want to create a measure that estimates the variance of a column in some table, for the entire population in the data.
Which DAX formula should you use ? [1]

A: VAR.S

√ B: VAR.P

C: VAR.Q

D: VAR.R

Q7:
Power BI comes with a wide range of visualizations to portray any data story you wish. Which Power BI visual is ideal for illustrating the fluctuations of stock prices over a given period of time? [1]

X A: Scatter chart

B: Waterfall chart

C: Treemap

D: Stacked column chart

Q8:
A retail analysis dashboard contains a small collection of datasets including Sales, Items and Stores. You have dragged the Category field from the Items dataset onto the report canvas. You now want to give users the option to select one, many or all Categories to filter the visuals on the report. Which element should you pick from the visualizations pane? [1]

√ A: Slicer

B: Multi-row card

C: Card

D: Funnel

Q9: Which relationship cardinality is not available in Power BI table relationships? [1]

A: One to Many

B: Many to One

√ C: Two to Many

D: One to One

Q10:
In the Power BI visualization pane, which field is not available when creating a Pie chart for visualization? [1]

A: Values

√ B: Group

C: Legend

D: Details

Student: Damian Vather

Programme: Data Science

Test: Hypothesis Testing 2020 :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Below Average Test performance
You got less than 80% for the Test:

Fundamentals : Explain : Hypothesis Testing 2020

Q1: A larger sample results in a more accurate mean and ensures a normal distribution. [1]

√ A: False

B: True

Q2:
There are 118 students who wrote the gather exam ,the students got an average of 57.60% ,following a normal distribution with a standard deviation of 27.60%.What is the confidence interval that 95% of exam marks will lie within? [1]

A: (52.57, 62.67)

√ B: (52.62, 62.58)

C: (51.62, 62.58)

D: (52.57, 62.58)

Q3:
If we used a t-distribution and used smaller degrees of freedom what will happen to the curves and the weight in the tails of the t-distribution curve?What would happen to a standard t-distribution if you decreased the degrees of freedom ? [1]

√ A: The peaks would flatten, tails would become larger

B: The peaks and tails would not change

C: The peaks would narrow, tails would become smaller

D: The peaks would widen, tails would become smaller

Q4: Which of the following is a property of the normal distribution? [1]

A: The total area under the curve can be anything between 0 and 1

B: 98.2% of the data falls within one standard deviation of the mean

C: 60% of the data points are to the left of the mean and 40% are to the right of mean

√ D: The mean, mode and median are all equal

Q5:
What distribution is used to test for equality of variances between two populations given that data was collected independently? [1]

A: t-distribution

B: Binomial distribution

C: Normal distribution

√ D: F-distribution

Q6:
A university wants to estimate the overall student average test score,taking into account 122 random student test scores.These selected test scores have an average of 68%
and a standard deviation of 14%.What is the confidence interval that will account for 95% of student averages? [1]

A: (61.34%, 72.66%)

B: (60.4%, 73.6%)

C: (65.18%, 68.82%)

√ D: (65.52%, 70.5%)

Q7: Which of the following statements is True? [1]

A: The null hypothesis is another term for the alternative hypothesis

B: A type 1 error is when the null hypothesis is rejected when it is in fact false

√ C: A type 2 error is when the null hypothesis is not rejected when it is in fact false

D: A type 1 error is when we fail to reject the null hypothesis when it is false

Q8:
The old iphone battery lasts 6 hours. A new iphone is advertised to last longer and we want to test this theory. We do this by charging and letting the phone battery die 30 times which takes 210 hours, with a standard deviation of 2 hours. How confident can we be that this new phone lasts longer than the original iphone? [1]

√ A: greater than 90%

B: 78%

C: less than 1%

D: 99%

Q9:

A food delivery service claims that they will deliver within 50 minutes.To test this claim, you place 41 food orders, timing each delivery.The average time taken by the delivery service is 53 minutes,with a standard deviation of 5 minutes.How confident are you that the food delivery service's claim is false? [1]

A: 75%

B: 90%

C: greater than 99%

√ D: less than 1%

Student: Damian Vather

Programme: Data Science

Test: Simple Linear Regression 2020 :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: In the equation below, what does q represent? $y = px + q$ [1]

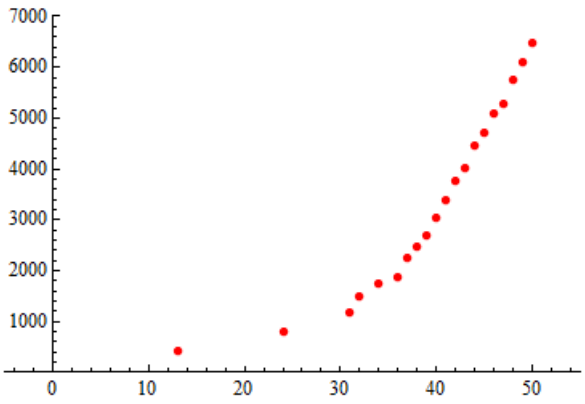
A: q is the x-intercept of the linear model.

√ B: q is the y-intercept of the linear model.

C: q is the slope, or gradient, of the linear model.

D: q is an unknown quantity.

Q2:
Refer to the figure below. A simple linear regression would be an appropriate method to model the data shown on this scatter plot.



[1]

A: True

√ B: False

Q3:
If we observe a point (3, 5.5), what is the residual (not the error) of this observation, with respect to the model below? $y = 2x + 3$ [1]

A: 9.0

B: -9.0

C: 3.5

√ D: -3.5

Q4: In the equation below, what does p represent? $y = px + q$ [1]

√ A: p is the slope, or gradient, of the linear model.

B: p is the y-intercept of the linear model.

C: p is the x-intercept of the linear model.

D: p is an unknown quantity.

Q5:
When we are assessing the accuracy of the model, the following method is a measure of the proportion of variance explained by the model. [1]

A: RSS (Residual Sum of Squares)

B: RMSE (Root Mean Square Error)

√ C: R^2 (R-Squared)

D: RSE (Residual Standard Error)

Q6: What is true about the gradient of the function below? $y = 2x + 3$ [1]

√ A: When the value of the x-variable is 0, the y-variable will be equal to 3.

B: When the value of the x-variable is 2, the y-variable will be equal to 0.

C: If we increase the x-variable by 2 units, the y-variable will increase by 1 unit.

D: The y-variable is always 3 units greater than the x-variable.

Q7: What is true about the slope of the function below? $y = 4x + 3$ [1]

√ A: For an increase of 1 unit in the x-variable, y-increases by 4 units.

B: For an increase of 2 units in the x-variable, y-increases by 1 unit.

C: The x-variable has a negative relationship with the y-variable.

D: When the value of the x-variable is 0, the y-variable will be equal to 2.

Q8: Least squares is a method of fitting a regression line which is robust (i.e: safe from) outliers. [1]

A: True

√ B: False

Student: Damian Vather

Programme: Data Science

Test: Model Selection and Metrics 2020 :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: What does RSS stand for? [1]

A: Regression Sum of Squares

B: Residual Squared State

C: Regression Squared State

√ D: Residual Sum of Squares

Q2:
Compare the three models below: Model 1: train MSE = 0.423, test MSE = 0.978 Model 2: train MSE = 0.572, test MSE = 0.644 Model 3: train MSE = 0.218, test MSE = 1.103 Based on this information, which of these models generalises the best to unseen data? [1]

A: Model 1

B: Model 3

C: There is no indication of one being better than the others

√ D: Model 2

Q3:
Compare the three models below: Model 1: train MSE = 0.423, test MSE = 0.978 Model 2: train MSE = 0.572, test MSE = 0.644 Model 3: train MSE = 0.218, test MSE = 1.103 Based on this information, which of these models generalises the worst to unseen data? [1]

A: Model 2

√ B: Model 3

C: Model 1

D: There is no indication of one being worse than the others

Q4: When adding more variables to a linear model, what is true about the R-squared value? [1]

A: R-squared cannot be calculated for models that use multiple predictor variables.

B: R-squared is not affected by adding additional variables.

C: Adding more variables will always decrease R-squared.

√ D: Adding more variables will always increase R-squared.

Q5: Adding more variables always leads to a lower test MSE. [1]

A: True

√ B: False

Q6: What does RMSE stand for? [1]

√ A: Root Mean Squared Error

B: Really Mean Statistical Error

C: Root Median Sum of Errors

D: Random Mean Sampling Error

Q7:
The R-squared measure is said to take on a 'proportion' of some attribute associated with the model. What is that proportion? [1]

√ A: Proportion of variance explained.

B: Proportion of outputs correctly predicted.

C: Proportion of predictor variables contributing to output.

D: Proportion of observations used for training.

Q8:
When our predictor variables have ranges and units that are quite different, it is pertinent to scale them before using them in a regression. Which of the following statements regarding scaling is FALSE? [1]

√ A: Standardisation is the process of squeezing a range of values to into the range [0,1].

B: Normalisation is the process of squeezing a range of values to into the range [0,1].

C: Normalisation is a scaling process which is inherently sensitive to outliers in the data.

D: Standardisation centres and scales a set of values such that they all have a mean of 0 and standard deviation of 1.

Q9:

Ridge and LASSO regression are known as regularisation techniques which we can use to improve a linear regression model. Which statement regarding these techniques is FALSE? [1]

A: Both methods shrink the magnitude of variable coefficients, but LASSO can shrink the values all the way to zero.

√ B: Both ridge and LASSO regression are able to shrink coefficients all the way to zero.

C: Ridge regression attempts to minimise the RSS as well as the L2-norm.

D: LASSO regression attempts to minimise the RSS as well as the L1-norm.

Student: Damian Vather

Programme: Data Science

Test: Regression MCQ exam 2020 :

Mark: 85.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1:
You want to measure the difference between the true y-value of each data point and the predicted value, which is the best method you could use? [1]

√ A: Residual Sum of Squares

B: The error method

C: Residual Mean error

D: Mean error method

Q2: The equation $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is used to calculate the: [1]

√ A: Mean Squared Error

B: R^2

C: Residual Sum of Squares

D: Logarithmic Residual Sum of Squares

Q3: Both β_0 and β_1 in the multiple linear regression equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$ are known as coefficients. [1]

A: True

√ B: False

Q4:
Which of the following is true regarding a multiple linear regression model trained on a dataset where all features have positive values?

(i) Coefficients can be both positive and negative

(ii) Coefficients are all equal

(iii) You will have zero test error

(iv) The resulting multiple linear regression model will have the general equation $y = \beta_0 + \beta_1 x_1$

[1]

A: (i) and (ii)

√ B: (i) only

C: (i), (ii) and (iv)

D: All of the above

Q5: Which of the following is False regarding Variables and Variable Selection? [1]

√ A: Method of Variance Thresholds is the same as The Correlation Method

B: Shrinkage methods can also be considered as a method of variable selection

C: If you ignore multicollinearity the model you are using is likely to have collinearity issues

D: Method of Variance is Easy and relatively safe way to reduce dimensionality at the start of the modeling process

Q6:
Which of the following statements are true? (i) *L2_norm* is known as the sum of the squares of the coefficients. (ii) *L1_norm* is known as the sum of the absolute values of the coefficients. (iii) We can use Shrinkage method in ridge regression to shrink coefficient's so that they are equal to zero (iv) In Ridge regression, we minimise RSS and the α(*L1_norm*) [1]

A: (i) and (ii)

B: (i) only

C: All of the above

X D: (i), (ii) and (iv)

Q7:
Variable selection and shrinkage (regularisation) are methods that can help with removing outliers in the data [1]

A: True

√ B: False

Q8: Standardisation is a method of feature scaling which is more robust to handling outliers. [1]

√ A: True

B: False

Q9:
LASSO regression can be considered as an implementation of which of the following: (i) Shrinkage methods (ii) Feature selection (iii) Scaling (iv) Normalisation [1]

A: (iii) only

B: (ii) and (iii)

C: (iv) only

√ D: (i) and (ii)

Q10:
Which of the following statements are true regarding Random Forest models?

(i) Random Forests are an example of a heterogeneous ensemble model

(ii) Random Forests are trained using the boosting method

(iii) Individual estimators are trained with different subsets of the data

(iv) Random Forests are generally more prone to overfitting compared to decision trees

[1]

A: (iii) only

B: (i) only

C: (i) only (iv)

X D: (i), (iii) and (iv)

Q11:
Practical Questions

Questions 11 - 20, are practical questions based on the given jupyter notebook file and dataset. Students are expected to fill in the missing code and use the resulting functions to answer the following questions.

What is the result of printing out the 6th column and the 13th row of X_train? [1]

A: -1.2508601954469347

B: -0.09303318078696134

√ C: -0.8282787380653501

D: 0.056380810844757615

Q12: What is the result of printing out 6th column and the 13th row of X_test? [1]

A: 1.7170736234873545

√ B: -0.17191842758365714

C: 0.5415166925051395

D: -1.5508940181797966

Q13: What is the result of printing out the 16th row y_train? [1]

A: 5

B: 4

√ C: 6

D: 7

Q14: What is the result of printing out the 16th row of y_test? [1]

A: 8

B: 7

√ C: 5

D: 6

Q15: What is the result of printing out model.intercept_ for the fitted model rounded to 3 decimal places? [1]

A: 5.800

B: 6.001

√ C: 5.821

D: -0.228

Q16: What is the result of printing out model.coef_[2] for the fitted model rounded to 2 decimal places? [1]

A: -0.46

B: -0.23

√ C: -0.26

D: -0.29

Q17: What is the Residual Sum of Squares value of the fitted LinearRegresion model on the testing set? [1]

√ A: 882.30

B: 883.30

C: 868.86

D: 725.94

Q18:
What is the Residual Sum of Squares value of the fitted Decision Tree Regression Model on the testing set? [1]

A: 1074.0

B: 1094.0

C: 4622.0

√ D: 1113.0

Q19: What is the result of printing out mean_abs_err(np.array([7.5,7,1.2]),np.array([3.2,2,-2])) ? [1]

√ A: 4.167

B: 3.367

C: 12.5

D: 4.65

Q20: Which regression model (LinearRegression vs DecisionTree) has the lowest Mean Absolute error? [1]

X A: LinearRegression

B: DecisionTree

Student: Damian Vather

Programme: Data Science

Test: Classification theory test 2020 :

Mark: 70.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: Which of the following statements is correct? [1]

√ A: Logistic Regression is a parametric model whereas a K-Nearest Neighbours model is non-parametric

B: The logistic regresssion model can detect and remove outliers in the data

C: The more data we have, the faster our logistic regression model trains

D: We can get more accurate predictions on a binary classification dataset by training a logistic regression model on each class individually and then adding the two model predictions together

Q2:
Assuming you have done the following: 1) imported LogisticRegression from sklearn.linear_model
2) imported train_test_split from sklearn.model_selection
3) cleaned and split your dataset to into X and y Which lines of code have an error? i) X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)
ii) model = LogisticRegression(C=0.1, penalty='l1')
iii) model.fit(X_train_y_train)
iv) pred = model.predict(y_test) [1]

A: i), iv)

B: iv)

√ C: iii), iv)

D: ii)

Q3: When building a logistic model we can do the following: [1]

A: Automatically optimise the C parameter in the LogisticRegression() object

B: Change the value of C to -1.00

X C: Assign a penalty of l2 and l1

D: Change the probability threshold from 0.5 to 0.6

Q4:
You built a model to detect COVID-19 cases in rural areas based on daily population movements. Your classifier had the following results: Correctly identified negative cases: 100241
Correctly identified positive cases: 95
Incorrectly identified negative cases: 24020
Incorrectly identified positive cases: 28 Which of the following metrics would best assess your model's performance? [1]

A: Precision

B: Accuracy

√ C: F1-Score

D: Recall

Q5:
You built a model which classifies if you're a python-wiz or a python-was. Calculate the F1-Score based on the following data: Correctly identified as a python-wiz : 126
Correctly identified as a python-was : 58
Incorrectly identified as a python-wiz: 14
Incorrectly identified as a python-was: 22 [1]

A: 0.89

X B: 0.84

C: 0.85

D: 0.88

Q6: Which one of the following statements is correct regarding the parameter C in logistic regression? [1]

A: C can only be used for binary logistic regression.

B: C can be assigned a number less than or equal to zero.

C: C is directly proportional to the regularisation coefficient, lambda.

√ D: An increase in C will result in less regularisation.

Q7: Select the correct statements from the following: i) Sensitivity refers to the True Positive Rate
ii) F1-Score is the harmonic mean between precision and recall
iii) "LogisticRegression" can be imported from "sklearn.linear_model"
iv) SVM and Random Forests can be used for classification [1]

A: i), iii)

B: i), ii), iii)

C: ii), iii)

√ D: All of the above

Q8: What will the following piece of code output?

```
from sklearn.metrics import classification_report
print('Classification Report')
print(classification_report(y_test, pred_lm, target_names=['ham', 'spam']))
```

[1]

A: A 2x2 confusion matrix for classifying "ham" or "spam".

√ B: A matrix of for assessing precision, recall, f1-score and support for classifying "ham" or "spam", along with micro, macro and weighted averages for the model.

C: A list of all the correct and incorrect classifications for "ham" or "spam" predicted by the model.

D: A table of accuracies for "spam" and "ham".

Q9: Which statement regarding the assessment of a classification model is true? [1]

√ A: The ROC curve can display both true positive and false positive rates at a range of thresholds.

B: AUC graphs do not show the overall performance of a classifier.

C: Precision is calculated as the the number of trup positives divided by the sum of true positives and true negatives.

D: Recall refers to the percentage of total incorrect classifications by a model.

Q10: Which of the following research scenarios could be solved using a classification model: [1]

A: Sorting an aerial photograph into urban and agricultural zones.

B: Fitting a generative model to a set of cat images and then generating a new cat image.

C: Predicting the number of litres of beer consumed after level 3 lockdown restrictions have been implemented.

X D: Predicting the time of rehabilitation for COVID-19 infected patients.

Student: Damian Vather

Programme: Data Science

Test: NLP theory test 2020 :

Mark: 90.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: Natural Language Processing is part of which field: i. Neuroscience
ii. Computer Science
iii. Artificial Intelligence
iv. Linguistics [1]

A: i, iii, iv

√ B: ii, iii, iv

C: All of the above

D: i, ii, iii

Q2: What is the main difference between stemming and lemmatization? [1]

√ A: Stemming can often create non-existent words, whereas lemmas are actual words

B: There is no difference, they are just different python libraries for doing the same thing

C: Stemming comes from the pandas python library while lemmatization comes from NLTK

D: Stem words can be looked up in a dictionary while lemmas cannot

Q3: Converting text into tokens and then converting them into integers or floats can be done using: [1]

X A: TF-IDF

B: Lemmatization

C: Bag of Words

D: CountVectorizer

Q4: Which one of the following are keyword normalization techniques? i. Stemming
ii. Part of Speech
iii. Named entity recognition
iv. Lemmatization [1]

√ A: i,iv

B: iii,ii

C: ii,iii,i

D: All of the above

Q5:
The process of identifying people, locations, or organizations from a given sentence or paragraph is called: [1]

√ A: Named entity recognition

B: Lemmatization

C: Stop word removal

D: Stemming

Q6:
Which of the following techniques can be used in the process of converting a keyword into its base form? [1]

A: Cosine Similarity

B: CountVectorizer

C: N-grams

√ D: Lemmatization

Q7:
N-grams are defined as the combination of N consecutive words. How many bi-grams can be generated from the sentence: "Classification is a great machine learning technique to learn"? [1]

A: 9

√ B: 8

C: 10

D: 7

Q8:
What would be the root word if we used the PorterStemmer() as the stemming algorithm on the word "loving"? [1]

A: lovi

B: Lovely

C: Lov

√ D: Love

Q9: When we want to separate each word in a sentece into a collection of tokens, we need to use: [1]

A: from nltk.tokenize import word_tokens

B: from nltk.stem import WordNetLemmatizer

√ C: from nltk.tokenize import word_tokenize

D: from nltk.stem import PorterStemmer

Q10:
In the sentence: "This is an introduction to Natural Language Processing", which words are considered stop words? [1]

√ A: This, is, an, to

B: Is , an

C: This, an, to

D: Is, an, to

Student: Damian Vather

Programme: Data Science

Test: Dimensionality Reduction MCQ :

Mark: 76.47 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Below Average Test performance
You got less than 80% for the Test:

Machine Learning : Unsupervised Learning : Dimensionality Reduction MCQ

We recommend that you do the following Train/Pre-processing:

Machine Learning : Regression : Variables and Variable Selection
Machine Learning : Regression : Fast.AI Lesson 5: Extrapolation and RF from Scratch
Machine Learning : Regression : Regularization - Ridge

Q1: Principal Component Analysis is a linear dimensionality reduction technique. [1]

A: False

√ B: True

Q2: Before performing PCA, what should ideally be done to the data set? [1]

A: encode variables

B: categorise variables

X C: normalise variables

D: standardize variables

Q3:
What can be said of the largest eigenvalue, in terms of its relation to the principal components of a dataset? [1]

A: it is equal to the amount of variance in the last principal component

√ B: it corresponds to the direction in which there is the largest amount of variance in the data

C: it is equal to the amount of variance in the first principal component

D: it corresponds to the direction in which there is the least amount of variance in the data

Q4: Which of the following is NOT a method for dimensionality reduction? [1]

√ A: k-means clustering

B: Principal Component Analysis

C: t-SNE

D: Multidimensional Scaling

Q5: t-SNE is a linear dimensionality reduction technique. [1]

A: True

√ B: False

Q6: Which of the following is not true regarding dimensionality reduction: [1]

X A: it decreases the interpretability of models

B: it can be used to aid data visualization

C: it decreases the computational time for training models

D: it always improves performance of clustering algorithms

Q7: t-SNE has one tunable hyperparameter, namely: [1]

A: irregularity

√ B: perplexity

C: complexity

D: duplicity

Q8: Which of the following dimensionality reduction techniques preserves distances between points? [1]

√ A: Multidimensional scaling

B: t-SNE

Q9: It is not necessary to have a target variable for applying dimensionality reduction algorithms. [1]

√ A: True

B: False

Q10:
The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA?
1. PCA is an unsupervised method
2. It searches for the directions that data have the largest variance
3. Maximum number of principal components <= number of features
4. All principal components are orthogonal to each other [1]

A: 1, 3 and 4

B: 1, 2 and 3

X C: 1, 2, and 4

D: 1, 2, 3 and 4

Q11:
In which of the following scenarios is t-SNE better to use than PCA for dimensionality reduction while working on a local machine with minimal computational power? [1]

A: Data set with 1 Million entries and 300 features

B: Data set with 100000 entries and 310 features

X C: Data set with 10,000 entries and 200 features

D: Data set with 10,000 entries and 8 features

Q12: In t-SNE algorithm, which of the following hyper parameters can be tuned? [1]

√ A: Number of dimensions, smooth measure and max iterations

B: Smooth measure of effective number of neighbours

C: Maximum number of iterations

D: Number of dimensions

Q13: Which of the following statement is correct for t-SNE and PCA? [1]

A: t-SNE is linear whereas PCA is non-linear

B: t-SNE and PCA both are linear

√ C: t-SNE is nonlinear whereas PCA is linear

D: t-SNE and PCA both are nonlinear

Q14: What will happen when eigenvalues are roughly equal? [1]

A: The number of dimensions in the data will increase

B: PCA will perform optimally

√ C: PCA will perform poorly

D: All eignevectors will be the same

Q15: An embedding is a representation of a vector in a different feature space. [1]

√ A: True

B: False

Q16:
In Multi-dimensional Scaling, a larger stress value is preferred, and it is this quantity that is maximised by MDS. [1]

A: True

√ B: False

Q17: Regarding an MDS scatter plot, which of the following is false: [1]

A: The axes themselves are meaningless

B: The orientation of the figure is arbitrary

√ C: The distances between observations are not preserved

D: he important thing is the proximity of the points

Student: Damian Vather

Programme: Data Science

Test: Clustering MCQ :

Mark: 69.57 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Below Average Test performance
You got less than 80% for the Test:

Machine Learning : Unsupervised Learning : Clustering MCQ

We recommend that you do the following Train/Pre-processing:

Machine Learning : Classification : Comet Starter Notebook
Machine Learning : Classification : Setting Up Comet
Machine Learning : Unsupervised Learning : Introduction to Recommender Systems

Q1: Which of the following is NOT a form of clustering? [1]

A: Linear Discriminant Analysis

B: Latent dirichlet allocation

X C: Dendrograms

D: Gaussian mixture modelling

Q2: The k-means clustering algorithm is sensitive to outliers. [1]

√ A: True

B: False

Q3:
When applying the k-means algorithm to a multi-dimensional data set, which of the following is a necessary pre-processing step? [1]

A: Normalize the data

B: Reduce the size of the data

√ C: Standardize the data

D: Remove a few features from the data

Q4: Which of the following is NOT true about the Partitioning Around Medoids (PAM) algorithm? [1]

X A: Any form of distance metric may be used

B: It is not necessary to standardize the data set before clustering

C: It is more robust to outliers than the k-means algorithm

D: The medoids are actual points in the data set

Q5: Which of the following is not a type of distance metric? [1]

A: Manhattan

√ B: Parisian

C: Minkowski

D: Euclidean

Q6: In Hierarchical clustering, there are two types of methods, these are:
i) Agglomerative
ii) Minimalist
iii) Divisive
iv) Partitioning [1]

√ A: i. and iii.

B: iii. and iv.

C: ii. and iii.

D: i. and ii.

Q7: What is the purpose of using the 'Elbow method'? [1]

√ A: It is used to determine the optimal number of clusters to use

B: It is used for cluster profiling

C: It helps us to determine which features are most important

D: It is used to measure the size of the clusters

Q8: Which values are typically chosen to be the X and Y axes of the Elbow plot, respectively? [1]

A: Number of features and average distance between clusters

√ B: Number of clusters and total within-cluster sum of squares

C: Number of clusters and average number of members

D: Number of clusters and average cluster size

Q9: The DBSCAN algorithm results in every point in a data set belonging to a cluster. [1]

A: False

X B: True

Q10: Which of the following machine learning problems would typically be solved using clustering? [1]

A: Screening of loan applications

B: Predicting the sale price of a recently listed house

√ C: Mining of customer insights to improve targeted marketing

D: Organising a person's phone camera images into labelled folders

Q11:
The k-means algorithm is a deterministic process (when run on the same data set, the cluster centroids will always take on the same values). [1]

√ A: False

B: True

Q12:
Consider the following two points in 2-dimensional space: A (1, 1), B (5, -3) What is the Euclidean distance between A and B? [1]

√ A: 5.66

B: 0

C: 32.00

D: 8.72

Q13: For the same two points as in Q12, what is the Manhattan distance between A and B? [1]

√ A: 8

B: -8

C: 4

D: 0

Q14: How does the value of k affect the computational complexity of the k-means algorithm? [1]

A: The complexity is unaffected by the value of k

B: An increase in k leads to a decrease in the complexity

X C: K can either increase or decrease the complexity depending on the distance metric chosen

D: An increase in k leads to an increase in the complexity

Q15: What is the minimum no. of variables/ features required to perform clustering? [1]

√ A: 1

B: 2

C: 3

D: 0

Q16:
When using the sklearn.cluster.KMeans function, what is the purpose of increasing the n_init argument? [1]

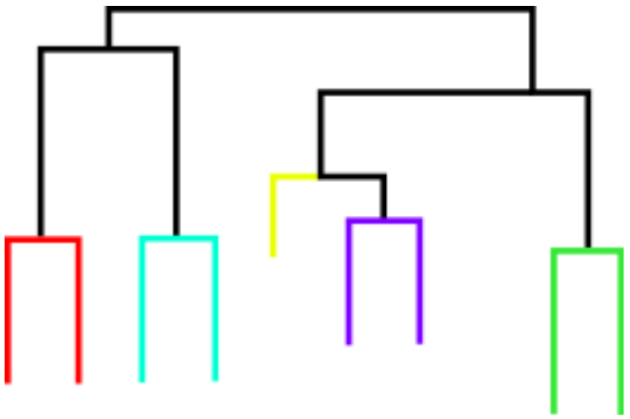
A: To speed up the algorithm

B: To specify the number of clusters

√ C: To avoid falling into local minima

D: To avoid overfitting

Q17:
What is the most appropriate number of clusters for the data represented by the following dendrogram:



[1]

X A: 2

B: 25

C: 4

D: 10

Q18: Which of the following is NOT a type of linkage method for hierarchical clustering? [1]

A: Single

√ B: Partial

C: Average

D: Complete

Q19: In which of the following cases will K-means clustering give poor results?

i) Data with outliers
ii) Non-linearly separable data (e.g. concentric clusters)
iii) Data with linearly separable clusters that don't overlap with each other
iv) Data with vastly different (i.e. different by a factor of 100) cluster sizes [1]

√ A: i,ii & iv

B: i & ii

C: ii & iii

D: i,ii,iii & iv

Q20: In Gaussian Mixture Model Clustering, the number of clusters K does not need to be specified. [1]

X A: True

B: False

Q21: Which statement is false regarding soft clustering? [1]

A: Soft clustering techniques are better suited for cases where clusters overlap with each other

B: Soft clustering techniques keep all possibilities of cluster assignment

√ C: Clusters are mutally exclusive

D: Data points can belong to multiple clusters

Q22:
In recommender systems, content-based methods are less likely to suffer from the cold start problem than collaborative approaches. [1]

X A: False

B: True

Q23: tf-idf is a method of which recommender system type? [1]

A: collaborative filtering

B: none of the above

C: hybrid

√ D: content-based

Student: Damian Vather

Programme: Data Science

Test: Unsupervised Exam :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Below Average Test performance
You got less than 80% for the Test:

Machine Learning : Unsupervised Learning : Unsupervised Exam

Q1:
The practical questions of this Exam should be answered using the attached UFO.csv file. This file contains information pertaining to various UFO sightings over the last 70 years. Use default parameters for the practical questions unless otherwise specified. Which of the following models is best suited to classify 10,000 rows of unlabelled Twitter data?

1. Naïve Bayes
2. Logistic regression
3. SVM
4. Linear regression

[1]

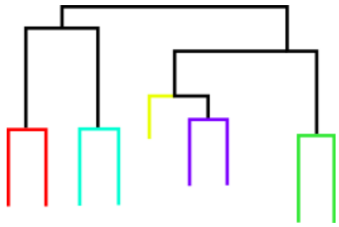
A: 1, 2, 3

√ B: None of these

C: 1, 2, 3, and 4

D: 1, 3

Q2: What is the appropriate choice for the number of clusters, given this dendrogram?



[1]

√ A: 4

B: 5

C: 3

D: 2

Q3:
Using K-means and the Elbow method, what is an appropriate number of geospatial clusters for this data? [3]

√ A: 4

B: 6

C: 5

D: 3

Q4:
Cluster the data geospatially using K-means, specifying 5 clusters. The largest cluster contains approximately what portion of the data? [2]

√ A: 40%

B: 30%

C: 35%

D: 50%

Q5:
If we plot these clusters geographically - which country appears to have the densest population of UFO sightings? [1]

A: New Zealand

B: Canada

C: UK

√ D: USA

Q6:
Plot a dendrogram of the years of UFO sightings using the ward method, using only the first 1000 entries. What is the optimal number of clusters judging by this dendrogram? [2]

√ A: 2

B: 5

C: 4

D: 3

Q7:
If we attempt to visualise a dendrogram of the years of UFO sightings of the full data set, what error will we most likely get? [1]

A: TypeError

B: ParseError

C: ValueError

√ D: MemoryError

Q8:
Using a TfidVectorizer (with English stopwords), what are the 3 “most important” *unique words* found in the comments column? [2]

A: sky, moving, craft

B: bright, light, moving

√ C: light, object, sky

D: bright, object, orange

Q9:
Using the following vectorizer:

vectorizer = TfidfVectorizer(max_features=20, stop_words='english'),

fit-transform it to the comments column. Then use PCA (with 10 components and a random_state of 1) to determine the percentage of the variance that the first two principal components explain.
[3]

A: 34%

B: 25%

√ C: 22%

D: 17%

Q10: Which of the following is not a feature selection/extraction technique? [1]

√ A: Pipelines

B: Term frequency-inverse document frequency

C: PCA

D: Multidimensional scaling

Q11: In K-means clustering, the k parameter does not need to be defined initially. [1]

A: True

√ B: False

Student: Damian Vather

Programme: Data Science

Test: Workplace Conduct MCQ :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1:

Feedback should always be

[1]

A:

Negative

B:

Positive

C:

Specific

√

D:

Vague

Q2: When offering feedback it is important to focus on

[1]

A:

√ The outcome of the behaviour or action

B:

Anything the person has ever done

C:

The time of day

D:

All the details

Q3: Feedback should be offered

[1]

A:

√ As soon after the action, provided that the provider is calm

B:

As long as possible after the action

C:

Any time, time makes no difference

D:

Instantaneously

Q4: Responding to feedback is best done

[1]

A:

Loud and angrily

B:

With tears

C:

Rationally and fairly

√

D:

With a counter-attack

Q5: Receiving feedback should be seen as

[1]

A:

A contribution towards self-development

√

B:

A pat on the back

C:

A personal attack

D:

Useless and a waste of time

Q6: A good business plan does not

[1]

A:

Demonstrate knowledge of competitors

B:

Provide a detailed market analysis

C:

Recognize how the law may affect its goals

D:

√ Exclude financial information

Q7: True or False: raising capital refers to ways to finance your business

[1]

A: False

√ B: True

Q8: Which of the following is not a major type of information system?

[1]

A:

Decision Support Systems

B:

√ Management Finance Systems

C:

Office Automation Systems

D:

Knowledge Management Systems

Q9: True or False: all financial transactions in businesses are the same

[1]

A: True

√ B: False

Q10: Which of the following is not a method of communication in a business

[1]

A:

Instant messaging

B:

Video-conferencing tools

C:

Keeping it to yourself

√

D:

Emails

Student: Damian Vather

Programme: Data Science

Test: Practitioners Meeting MCQ :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1:

Which of the following is NOT TRUE of preparation for a meeting

[1]

A:

The agenda needs to be distributed in advance of the meeting

B:

The agenda and relevant timing needs to be planned in advance

C:

Practical arrangements need to be organised in advance

D:

The chairperson must be the last to arrive

√

Q2:

Which of the following is NOT a role of the chairperson of a meeting

[1]

A:

Provide snacks

√

B:

Set the agenda

C:

Maintain order

D:

Ensure fairness and equality

Q3: When does the chairperson’s role cease with regards to a meeting

[1]

A:

When the meeting begins

B:

When the first person departs

C:

Once the minutes have been written up, checked and distributed

√

D:

Whenever they feel like it

Q4: What is the chairperson’s primary role in a practitioner’s meeting

[1]

A:

To ensure time is not wasted

B:

To guide the meeting and steer members to work harmoniously and purposefully as a team

√

C:

To criticise work poorly done

D:

To do all the administration

Q5: Which of the following is something that the chairperson should not do during a practitioner's meeting?

[1]

A:

Criticise

√

B:

Clarify

C:

Communicate

D:

Control

Damian Vather

Explore Data Science Academy

Formative Evidence

ITEM	RESULT
Coding Challenge 1: Basic List Exercises	100%
Coding Challenge 2: Basic Sorting and List Comprehensions	100%
Coding Challenge 3: List Exercises	100%
Coding Challenge 4: Advanced List Exercises	92%
Regression Coding challenge 1 - 2020	100%
Regression Coding Challenge 3 - 2020	100%
Classification Coding Challenge 1 (2020) - Logistic Regression	100%
TOTAL	**98.86%**

Student: Damian Vather

Programme: Data Science

Test: Coding Challenge 1: Basic List Exercises :

Mark: 100.00 %

```
28-02-20_11:06:29
========================================
========== AUTOGRADING ================
========================================
Examining contents of 3739_1407.zip
One jupyter notebook found
----------------------------------------
Unzipping 3739_1407.zip to /home/autograder/unzipped
Archive: /home/autograder/temp/3739_1407.zip
inflating: /home/autograder/unzipped/3739_1407/coding_challenge_1_test_suite.py
inflating: /home/autograder/unzipped/3739_1407/coding_challenge_1_student_version-1407-1407.ipynb
Unzipped successfully
----------------------------------------
Processing notebook
Notebook processed successfully
----------------------------------------
Checking submission for syntax errors
Check complete
----------------------------------------
Grading student code
Student code graded successfully
PASSED return_type_is_list
PASSED front_x
PASSED front_x
PASSED front_x
PASSED front_x_unseen
PASSED front_x_unseen
PASSED rna_length
PASSED rna_length
PASSED rna_length
PASSED rna_length
PASSED rna_length_unseen
PASSED rna_length_unseen
PASSED rna_length_unseen
PASSED rna_length_unseen
----------------------------------------
Saving result (100.0) to 3739_1407.csv
========================================
AUTOGRADING COMPLETED WITH EXIT STATUS: 0
========================================

Submission Name : 3739_1407.zip
Grade : 100.0
```

Basic list exercises

This notebook is a quick assessment of your progress on learning Python.

Instructions to Students

- **Do not add or remove cells in this notebook. Do not edit or remove the `### START FUNCTION` or `### END FUNCTION` comments. Do not add any code outside of the functions you are required to edit. Doing any of this will lead to a mark of 0%!**
- Answer the questions according to the specifications provided.
- Use the given cell in each question to to see if your function matches the expected outputs.
- Do not hard-code answers to the questions.
- The use of stackoverflow, google, and other online tools are permitted. However, copying fellow student's code is not permissible and is considered a breach of the Honour code below. Doing this will result in a mark of 0%.
- Good luck, and may the force be with you!

Honour Code

I **YOUR NAME, YOUR SURNAME**, confirm - by submitting this document - that the solutions in this notebook are a result of my own work and that I abide by the EDSA Student Manifesto.

Non-compliance with the honour code constitutes a material breach of contract.

Question 1

Given a list of lowercase strings, return a list with the strings sorted in alphabetical order, except group all the strings that begin with `x` at the beginning of the list.
e.g. ['mix', 'xyz', 'apple', 'xanadu', 'aardvark'] yields
`['xanadu', 'xyz', 'aardvark', 'apple', 'mix']`

Hint: This can be done by making 2 lists and sorting each of them before combining them.

Hint: Remember that Python's `list` object has a `sort()` function. Python also has a built-in function called `sorted()`.

In [7]:

```
### START FUNCTION
def front_x(words):
    # your code here
    starting_x_list = [x for x in words if x.startswith('x')]
    rest_list = [x for x in words if x not in starting_x_list]
    return sorted(starting_x_list) + sorted(rest_list)
### END FUNCTION
```

You may upload to the Autograder as many times as you wish; It can be a useful ally.

If your output to the following function calls matches the corresponding output, then your solution may be working.

In [10]:

```
front_x(['jhsdfi', 'jeiri', 'iopqwu', 'bhvaiau', 'loaks'])
```

Out[10]:

['bhvaiau', 'iopqwu', 'jeiri', 'jhsdfi', 'loaks']

Expected Outputs:

```
front_x(['mix', 'xyz', 'apple', 'xanadu', 'aardvark']) == ['xanadu', 'xyz', 'aardvark', 'apple', 'mix']
front_x(['netowrk', 'artist', 'xamarian', 'king', 'cat']) == ['xamarian', 'artist', 'cat', 'king', 'netowrk']
front_x(['jhsdfi', 'jeiri', 'iopqwu', 'bhvaiau', 'loaks']) == ['bhvaiau', 'iopqwu', 'jeiri', 'jhsdfi', 'loaks']
```

Question 2

RNA Length

Proteins are the building blocks of life. The process starts when another protein (a polymerase) reads your genetic code and performs transcription, the code (RNA) for making protein. Ribosomes will then attach to this RNA code to begin making specific proteins. Your genetic code is made up of pairs of nucleotides. Once transcribed into the code, these nucleotides contain the following nucleotides:

U,G,A,C.

UGA, UAA, and UAG are stop codons and AUG is a start codon so that the protein, ribosome, can identify where to stop and start the protein formation.

A genetics lab has tasked you to design a function which recognises the length of the RNA code (excluding the start and stop codons). The function must also determine if their code is actually valid (must contain AUG as a start codon and UGA, UAA, or UAG at the end)

Note: RNA starts with a start codon AUG and ends with UGA, UAA, or UAG

The function must return "Not readable RNA code" if the above conditions are not met.

In [12]:

```
### START FUNCTION
def rna_length(mrna):
    # your code here
    if mrna.startswith('AUG') and (mrna[-3:].endswith('UGA') or mrna[-3:].endswith('UAA') or mrna[-3:].endswith('UAG')):
        return len(mrna[3:-3])
    else:
        return 'Not readable RNA code'
### END FUNCTION
```

In [16]:

```
rna_length('AUGUAGGCACAUUUAUGCUCCUGA')
```

Out[16]:

18

Expected Outputs:

```
rna_length('AUGUAGCAUAA') == 5
rna_length('AUGUUAUAG') == 3
rna_length('AUGUAGGCACAUUUAUGCUCCUGA') == 18
rna_length('AUGAGGCACCUUCUGCUCCUUAC') == "Not readable RNA code"
```

References

This tutorial was created using the wonderful Google Python Developers course, and has been modified under the Creative Common's licence 2.5

Google's Python Class http://code.google.com/edu/languages/google-python-class/

Student: Damian Vather

Programme: Data Science

Test: Coding Challenge 2: Basic Sorting and List Comprehensions :

Mark: 100.00 %

```
01-03-20_21:14:45
========================================
============ AUTOGRADING ================
========================================
Examining contents of 3739_1416.zip
One jupyter notebook found
----------------------------------------
Unzipping 3739_1416.zip to /home/autograder/unzipped
Archive: /home/autograder/temp/3739_1416.zip
inflating: /home/autograder/unzipped/3739_1416/coding_challenge_2_student_version-1416.ipynb
inflating: /home/autograder/unzipped/3739_1416/coding_challenge_2_test_suite.py
Unzipped successfully
----------------------------------------
Processing notebook
Notebook processed successfully
----------------------------------------
Checking submission for syntax errors
Check complete
----------------------------------------
Grading student code
Student code graded successfully
PASSED return_type
PASSED sort_last
PASSED sort_last
PASSED sort_last
PASSED sort_last
PASSED sort_last
PASSED sort_last_unseen
PASSED sort_last_unseen
PASSED sq_cube
PASSED sq_cube
PASSED sq_cube
PASSED sq_cube
PASSED sq_cube_unseen
PASSED sq_cube_unseen
PASSED dna_complementary
PASSED dna_complementary
PASSED dna_complementary
PASSED dna_complementary
PASSED dna_complementary_unseen
PASSED dna_complementary_unseen
----------------------------------------
Saving result (100.0) to 3739_1416.csv
========================================
AUTOGRADING COMPLETED WITH EXIT STATUS: 0
========================================

Submission Name : 3739_1416.zip
Grade : 100.0
```

Student: Damian Vather

Programme: Data Science

Test: Coding Challenge 3: List Exercises :

Mark: 100.00 %

```
08-03-20_21:50:57
========================================
========== AUTOGRADING ================
========================================
Examining contents of 3739_1418.zip
One jupyter notebook found
----------------------------------------
Unzipping 3739_1418.zip to /home/autograder/unzipped
Archive: /home/autograder/temp/3739_1418.zip
inflating: /home/autograder/unzipped/3739_1418/coding_challenge_3_test_suite.py
inflating: /home/autograder/unzipped/3739_1418/coding_challenge_3_student_version-1418.ipynb
Unzipped successfully
----------------------------------------
Processing notebook
Notebook processed successfully
----------------------------------------
Checking submission for syntax errors
Check complete
----------------------------------------
Grading student code
Student code graded successfully
PASSED remove_adjacent
PASSED remove_adjacent
PASSED remove_adjacent
PASSED remove_adjacent
PASSED remove_adjacent
PASSED remove_adjacent_unseen
PASSED remove_adjacent_unseen
PASSED square_odd
PASSED square_odd
PASSED square_odd
PASSED square_odd
PASSED square_odd_unseen
PASSED square_odd_unseen
PASSED symmetrical_sum
PASSED symmetrical_sum
PASSED symmetrical_sum
PASSED symmetrical_sum
PASSED symmetrical_sum_unseen
PASSED symmetrical_sum_unseen
----------------------------------------
Saving result (100.0) to 3739_1418.csv
========================================
AUTOGRADING COMPLETED WITH EXIT STATUS: 0
========================================

Submission Name : 3739_1418.zip
Grade : 100.0
```

Student: Damian Vather

Programme: Data Science

Test: Coding Challenge 4: Advanced List Exercises :

Mark: 92.31 %

```
22-03-20_21:53:25
========================================
========== AUTOGRADING ================
========================================
Examining contents of 3739_1420.zip
One jupyter notebook found
----------------------------------------
Unzipping 3739_1420.zip to /home/autograder/unzipped
Archive: /home/autograder/temp/3739_1420.zip
inflating: /home/autograder/unzipped/3739_1420/coding_challenge_4_student_version-1420.ipynb
inflating: /home/autograder/unzipped/3739_1420/coding_challenge_4_test_suite.py
Unzipped successfully
----------------------------------------
Processing notebook
Notebook processed successfully
----------------------------------------
Checking submission for syntax errors
Check complete
----------------------------------------
Grading student code
Student code graded successfully
PASSED linear_merge
PASSED linear_merge
PASSED linear_merge
PASSED linear_merge
PASSED linear_merge
PASSED linear_merge_unseen
PASSED linear_merge_unseen
PASSED linear_merge_unseen
PASSED linear_merge_unseen
PASSED linear_merge_unseen
FAILED linear_merge_is_linear
Inputs: [[3, 2, 1], [5, 6, 1]]
assert [1, 1, 2, 3, 5, 6] == [3, 2, 1, 5, 6, 1]
At index 0 diff: 1 != 3
Full diff:
your output: [1, 1, 2, 3, 5, 6]
expected output: [3, 2, 1, 5, 6, 1]
PASSED amino_acids
PASSED amino_acids
----------------------------------------
Saving result (92.31) to 3739_1420.csv
========================================
AUTOGRADING COMPLETED WITH EXIT STATUS: 0
========================================

Submission Name : 3739_1420.zip
Grade : 92.31
```

Student: Damian Vather

Programme: Data Science

Test: Regression Coding challenge 1 - 2020 :

Mark: 100.00 %

```
03-08-20_19:32:09
========================================
============ AUTOGRADING ================
========================================
Examining contents of 3739_1696.zip
One jupyter notebook found
----------------------------------------
Unzipping 3739_1696.zip to /home/autograder/unzipped
Archive: /home/autograder/temp/3739_1696.zip
inflating: /home/autograder/unzipped/3739_1696/regression_challenge_1_student_version-1696.ipynb
inflating: /home/autograder/unzipped/3739_1696/regression_challenge_1_test_suite.py
inflating: /home/autograder/unzipped/3739_1696/preamble.py
Unzipped successfully
----------------------------------------
Processing notebook
Notebook processed successfully
----------------------------------------
Checking submission for syntax errors
Check complete
----------------------------------------
Grading student code
Student code graded successfully
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
```

```
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_1
PASSED question_2
PASSED question_2
PASSED question_2
PASSED question_2
PASSED question_2
PASSED question_2
PASSED question_2
PASSED question_2
PASSED question_2_ValueError
PASSED question_2_ValueError
PASSED question_3
PASSED question_3
PASSED question_3
PASSED question_3
PASSED question_3
PASSED question_3
----------------------------------------
Saving result (100.0) to 3739_1696.csv
========================================
AUTOGRADING COMPLETED WITH EXIT STATUS: 0
========================================

Submission Name : 3739_1696.zip
Grade : 100.0
```

Student: Damian Vather

Programme: Data Science

Test: Regression Coding Challenge 3 - 2020 :

Mark: 100.00 %

```
22-08-20_20:49:05
========================================
========== AUTOGRADING ================
========================================
Examining contents of 3739_1700.zip
One jupyter notebook found
----------------------------------------
Unzipping 3739_1700.zip to /home/autograder/unzipped
Archive: /home/autograder/temp/3739_1700.zip
inflating: /home/autograder/unzipped/3739_1700/regression_challenge_3_student_version_1700_(5)
(1).ipynb
inflating: /home/autograder/unzipped/3739_1700/regression_challenge_3_test_suite.py
inflating: /home/autograder/unzipped/3739_1700/preamble.py
Unzipped successfully
----------------------------------------
Processing notebook
Notebook processed successfully
----------------------------------------
Checking submission for syntax errors
Check complete
----------------------------------------
Grading student code
Student code graded successfully
PASSED question_1
PASSED question_1
PASSED question_1_shape
PASSED question_1_shape
PASSED question_2_X_train
PASSED question_2_X_train
PASSED question_2_y_train
PASSED question_2_y_train
PASSED question_2_X_test
PASSED question_2_X_test
PASSED question_2_y_test
PASSED question_2_y_test
PASSED question_3_type
PASSED question_3_predictions
PASSED question_3_predictions
PASSED question_4
PASSED question_4
----------------------------------------
Saving result (100.0) to 3739_1700.csv
========================================
AUTOGRADING COMPLETED WITH EXIT STATUS: 0
========================================

Submission Name : 3739_1700.zip
Grade : 100.0
```

Student: Damian Vather

Programme: Data Science

Test: Classification Coding Challenge 1 (2020) - Logistic Regression :

Mark: 100.00 %

```
27-09-20_21:04:34
========================================
========== AUTOGRADING ================
========================================
Examining contents of 3739_1794.zip
One jupyter notebook found
----------------------------------------
Unzipping 3739_1794.zip to /home/autograder/unzipped
Archive: /home/autograder/temp/3739_1794.zip
inflating: /home/autograder/unzipped/3739_1794/logistic_regression_test_suite.py
inflating:
/home/autograder/unzipped/3739_1794/Logistic_Regression_Challenge_student_version_1794_1794_Damo.ipynb
inflating: /home/autograder/unzipped/3739_1794/preamble.py
Unzipped successfully
----------------------------------------
Processing notebook
Notebook processed successfully
----------------------------------------
Checking submission for syntax errors
Check complete
----------------------------------------
Grading student code
Student code graded successfully
PASSED question_1_
PASSED question_1_
PASSED question_1_
PASSED question_2_X_train
PASSED question_2_y_train
PASSED question_2_X_test
PASSED question_2_y_test
PASSED question_3_1_type
PASSED question_3_1_intercept
PASSED question_3_1_coef
PASSED question_3_2
PASSED question_3_2
PASSED question_3_3_accuracy
PASSED question_3_3_accuracy
PASSED question_3_3_precision
PASSED question_3_3_precision
PASSED question_3_3_recall
PASSED question_3_3_recall
PASSED question_3_3_f1
PASSED question_3_3_f1
----------------------------------------
Saving result (100.0) to 3739_1794.csv
========================================
AUTOGRADING COMPLETED WITH EXIT STATUS: 0
========================================

Submission Name : 3739_1794.zip
Grade : 100.0
```

Damian Vather

Explore Data Science Academy

Summative Evidence

ITEM	RESULT
Analyse Theory Exam	100%
Gather Exam	100%
EDSA Final Exam - Part 1	91%
EDSA Final Exam - Part 2	75%
Supplementary Supervised - Exam	No Submission
Supplementary Analyse Exam	No Submission
TOTAL	**91.00%**

Student: Damian Vather

Programme: Data Science

Test: Analyse Theory Exam :

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: Which of the following are true regarding multiple statements per line in Python: [1]

A: Only variable assignment statements may occur multiply on a single line.

B: Placing multiple statements on a single line is discouraged by PEP 257.

√ C: Placing multiple statements on a single line is discouraged by PEP 8.

D: Multiple statements on the same line are separated by the & character.

Q2: Which of the following are true about lambda functions? [1]

A: Lambda, by its very nature, is not used to write simple functions without the use of defining them beforehand.

√ B: They make code more readable, where the purpose of the function is basically defined in one line of code.

C: The lambda function itself can be used to iterate through a list.

D: They are more efficient, where the function is basically not interpretable.

Q3: In Python, strings are… [1]

√ A: str objects

B: changeable

C: a pointer

D: mutable

Q4:
Which of the following would separate a string input_string on the first 2 occurrences of the letter “e”? [1]

A:

split(input_string = 'e', 2)

B:

split(input_string == 'e', 2)

C:

'e'.split(input_string, 2)

D:

input_string.split('e', maxsplit=2)

√

Q5: Python strings have a property called “immutability.” What does this mean? [1]

√ A: You cannot change an existing string. Otherwise create a new string that is a variation on the original.

B: You can update a string in Python with concatenation

C: Strings in Python can be represented as arrays of chars

D: Strings can’t be divided by numbers

Q6: What is the output for?

```
S = [['him', 'sell'], [90, 28, 43]]
S[0][1][1]
```

[1]

A:

'e'

√

B:

'h'

C:

28

D:

'i'

Q7: What is Instantiation in terms of OOP terminology? [1]

√ A: Creating an instance of class

B: Modifying an instance of class

C: Deleting an instance of class

D: Copying an instance of class

Q8: Which of the following is not a python built-in module? [1]

√ A: pandas

B: sys

C: random

D: math

Q9:
The time complexity of a quick sort algorithm which makes use of median, found by an O(n) algorithm, as pivot element is: [1]

√ A: *O(nlogn)*

B: *O(n)*

C: *O(nloglogn)*

D: $O(n^2)$

Q10: Which of these is true about recursion? [1]

A: It is not possible to write an iterative version of any recursive algorithm.

√ B: Recursion occurs when something is defined in terms of itself or of its type.

C: Recursive functions run faster than non-recursive function

D: Recursive functions are easy to debug.

Q11: How would you add changes to a repository on github inside a bash terminal? [1]

A:

√ git push

B:

git merge

C:

git add

D:

git upload

Q12: Complete the following piece of code, where "___" represents missing code:

```
def dice(num_of_rolls, seed):
    rolled_list = []
    random.___(seed)

    for i in range(0,num_of_rolls+1):
        rolled_list.append(___.randint(1,6))
        dict_of_rolls = ___ i : rolled_list[i] for i in range(1, len(rolled_list) ) ___

    return dict_of_rolls
```

And when called:

dice(5,42)

The following output is given:

>>> {1: 1, 2: 1, 3: 6, 4: 3, 5: 2}

[1]

A:

√ seed, random, { , }

B:

randint, np, [,]

C:

seed, np, [,]

D:

randint, random, { , }

Q13: Given the following python code, what would the output of the code give?

```
my_tuppy = (1,2,5,8)
my_tuppy[2] = 6
```

[1]

A: No output would be given as the tuple was not called

√ B: A TypeError

C: A ValueError

D: my_tuppy(1,6,5,8)

Q14:
After you create a new Git repository and create a file named edsa-boss.html, which of the following commands will not work if used in bash? [1]

A:

git add edsa-boss.html

B:

git add .

C:

git status

D:

√ git commit -m "git edsa boss web file added"

Q15: Which of these symbols typically represent the unbiased variance of a population: [1]

A: s

B: sigma

√ C: $sigma^2$

D: s^2

Q16: You toss a coin thrice. What is the probability of getting: i) at least two tails?
ii) No heads [1]

√ A: 1/2, 1/8

B: 1/4, 1/4

C: 1/3, 1/8

D: 1/2, 1/2

Q17: Regarding the central limit theorem, which one of the following statements is correct? [1]

A: The set of the sample means are not approximated to a normal distribution

B: The theorem approximates the sample means around an exponential distribution

√ C: A sample size of 30 and greater is a suggested size to use when applying calculations such as the mean and standard deviation, when the data is symmetric

D: The sample standard deviation increases as sample size increases

Student: Damian Vather

Programme: Data Science

Mark: 100.00 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1: Who won the Oscar for “Actor in a Leading Role” in 2015 ?
(Hint winner is indicated as '1.0') [1]

A: Natalie Portman

√ B: Leonardo DiCaprio

C: Eddie Redmayne

D: Michael Fassbender

Q2: What query will produce the ten oldest movies in the database? [1]

A:

```
SELECT TOP(10) * FROM movies WHERE release_date ORDER BY release_date ASC
```

B:

```
SELECT TOP(10) * FROM movies WHERE release_date IS NULL ORDER BY release_date DESC
```

C:

```
SELECT TOP(10) * FROM movies WHERE release_date IS NOT NULL ORDER BY release_date DESC
```

D:

```
SELECT TOP(10) * FROM movies WHERE release_date IS NOT NULL ORDER BY release_date ASC
```

√

Q3: What was the budget of the 5th most expensive movie in the database? [1]

A: 220 000 000

B: 380 000 000

C: 270000000

√ D: 260000000

Q4: How many unique awards are there in the Oscars table? [1]

A: 80

√ B: 114

C: 141

D: 53

Q5: How many movies in the database contain the word “Spider”? [1]

A: 5

B: 1

C: 0

√ D: 9

Q6: How many movies were released between 1 August 2006 and 1 October 2009 that have a popularity score of more than 40 and a budget of less 50 000 000? [1]

A: 23

√ B: 29

C: 18

D: 35

Q7: How many unique characters has Vin Diesel played so far in the database? [1]

√ A: 16

B: 19

C: 18

D: 24

Q8: What are the Genres of the movie “The Royal Tenenbaums”? [1]

√ A: Drama, Comedy

B: Crime, Thriller

C: Action, Romance

D: Drama, Romance

Q9: Which of the following movies has the keyword spy and also has the United Kingdom as a production country? [1]

A: The Naked Gun

√ B: Johnny English

C: The Man from U.N.C.L.E.

D: Casino Royale

Q10: What was the total amount of money spent to produce the movies that won the Best Picture award from 1990 up to and including 2010? [1]

A: 1 346 400 000

B: 954 500 000

√ C: 969 500 000

D: 2 340 900 000

Q11: What are the 3 production companies with the highest movie popularity score on average? [1]

√ A: The Donners' Company, Bulletproof Cupid and Kinberg Genre

B: Bulletproof Cupid, The Donners' Company ,MCL Films S.A

C: B.Sting Entertainment, Illumination Pictures and Aztec Musique

D: MCL Films S.A., Turner Pictures and George Stevens Productions

Q12: How many female actors (i.e. gender = 1) have a name that starts with N? [1]

√ A: 355

B: 7335

C: 0

D: 1949

Q13: Which of the genres has movies with the lowest average popularity score? [1]

A: Animation

√ B: Foreign

C: Documentary

D: Science Fiction

Q14: Which award category has the highest number of actor (all genders) nominations? (Hint Oscars.name contains both actors names and film names) [1]

√ A: Actor in a Supporting Role

B: Best Picture

C: Actress in a Supporting Role

D: Special Achievement Award

Q15: Which country has the 5th highest number of movie productions? [1]

A: Australia

B: China

C: Spain

√ D: Canada

Q16: For all of the entries in the Oscars table before 1934, the year is stored differently than in all the subsequent years. E.g the year would be saved as “1932/1933” instead of ju

A:

UPDATE Oscars SET year = SELECT LEFT(year, 4)

B:

UPDATE Oscars year = RIGHT(year, 4)

C:

√ UPDATE Oscars SET year = RIGHT(year, 4)

D:

UPDATE Oscars SET year = LEFT(year, 4)

Q17: DStv will be having a special week dedicated to Alan Rickman. Which of the following queries would create a new view that shows the titles, release dates, taglines and ove

A:

CREATE NEW VIEW Name = Alan_Rickman_Movies AS SELECT title, release_date, tagline, overview FROM Movies LEFT JOIN Casts ON Casts.movie_id = Movies.movie_

B:

VIEW Alan_Rickman_Movies AS SELECT title, release_date, tagline, overview FROM Movies LEFT JOIN Casts ON Casts.movie_id = Movies.movie_id Left JOIN Actors ON C

C:

√ CREATE VIEW Alan_Rickman_Movies AS SELECT title, release_date, tagline, overview FROM Movies LEFT JOIN Casts ON Casts.movie_id = Movies.movie_id Left JOIN Ac

D:

SELECT title, release_date, tagline, overview FROM Movies LEFT JOIN Casts ON Casts.movie_id = Movies.movie_id Left JOIN Actors ON Casts.actor_id = Actors.actor_id WH

Q18:
Which of the statements about database normalisation is true? i). Database normalisation removes inconsistencies which may cause the analysis of our data to be more complica efficiency, maintains data integrity and reduces the need to re-design the database if new data is introduced. iv). Database normalisation supports up to 3rd normal form and remov

A: (i),(ii)

√ B: (i),(iii)

C: (iii), (iv)

D: (ii), (iv)

Q19: Which of the following statements about Views is correct? [1]

A: Database views are virtual tables which are stored in the cloud and provide easy access to data

B: Database views are a result of a set of queries applied to the database.

√ C: Database views are virtual tables which are stored as objects and provide restricted access to data.

D: Database views are physically stored tables which provide restricted access to data.

Q20: Select the most correct statement about the Amazon RDS. (Hint: Use google if you're unsure) [1]

√ A: It is an AWS service used for hosting databases as instances.

B: It is a storage bucket.

C: It is an AWS service used for pipeline version control instances.

D: It is an AWS service used for storing csv files as databases.

Q21: Which of the following statements is true about AWS cloud services? (Hint: Use google if you're unsure)
i) You are responsible for the maintenance of the hardware and software components of the server. ii) AWS cloud services can be a free or a pay as you use service. iii) AV

A: (i),(ii) only and (iv).

B: (i) and (iv) only.

C: None of the above.

√ D: (ii) and (iii) only.

Q22: Let us say you would like to create a new table called “Watched” that contains all of the movies that you have already watched. This table should have only four columns, m

A:

NEW TABLE Watched (movie_id int NOT NULL PRIMARY KEY, title varchar(500), runtime float null, vote_average float null)

B:

CREATE DATABASE Watched (movie_id int NOT NULL PRIMARY KEY, title varchar(500), runtime float null, vote_average float null)

C:

NEW DATABASE Watched (movie_id int NOT NULL PRIMARY KEY, title varchar(500), runtime float null, vote_average float null)

D:

√ CREATE TABLE Watched (movie_id int NOT NULL PRIMARY KEY, title varchar(500), runtime float null, vote_average float null)

Q23: You changed your mind and now also want to include the release date in your new “Watched” table. What query would you use to make this change? [1]

A:

UPDATE TABLE Watched ADD release_date datetime

B:

ALTER TABLE Watched COLUMN(release_date datetime)

C:

√ ALTER TABLE Watched ADD release_date datetime

D:

UPDATE TABLE Watched COLUMN(release_date datetime)

Q24: Now we have to go about adding all of the movies that we have already watched. You are a very big Comedy movie fan and you know that you have watched every single C

A:

√ INSERT INTO Watched SELECT Movies.movie_id, Movies.title, Movies.runtime, Movies.vote_average, Movies.release_date FROM Movies LEFT JOIN GenreMap ON Movies.n

B:

SELECT INTO Watched Movies.movie_id, Movies.title, Movies.runtime, Movies.vote_average, Movies.release_date FROM Movies LEFT JOIN GenreMap ON Movies.movie_id =

C:

INSERT INTO Watched COLUMNS(movie_id, title, runtime, vote_average, release date) SELECT Movies.movie_id, Movies.title, Movies.runtime, Movies.vote_average, Movies.

D:

INSERT (SELECT Movies.movie_id, Movies.title, Movies.runtime, Movies.vote_average, Movies.release_date FROM Movies LEFT JOIN GenreMap ON Movies.movie_id = Ger

Student: Damian Vather

Programme: Data Science

Test: EDSA Final Exam - Part 1 :

Mark: 90.70 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1:
Gather Section (20 Questions, 40 Marks) You should have access to file called tmdb_5000.bak. The first step to answering this set of questions will be to restore this bak file to access the data. Based on that data, answer the following questions: What is the primary key for the table “movies”? (1 mark) [1]

A: title

B: movie_key

√ C: movie_id

D: film_id

Q2: How many foreign keys does the “languagemap” table have? (1 mark) [1]

A: 1

B: 0

√ C: 2

D: 3

Q3:
If the foreign keys in the “GenreMap” table is set to ON DELETE CASCADE, which of the following is true? (2 marks) [2]

A: Deleting an entry in “GenreMap” will lead to no other changes in the data.

B: Deleting an entry in the “Genre” table will lead to all entries with a matching “genre_id” in the “GenreMap” table to also be deleted.

C: Deleting an entry in the “GenreMap” table will lead to all entries with a matching “genre_id” in the “GenreMap” table to also be deleted.

X D: Deleting an entry in the “GenreMap” table will lead to all entries with a matching “genre_id” in the “Genre” table to also be deleted.

Q4: What code would you use to set up a view of all movies that did not get released? (1 mark) [1]

A: CREATE VIEW Not_Released SELECT * FROM movies WHERE release_status != 'Released'

√ B: CREATE VIEW Not_Released AS SELECT * FROM movies WHERE release_status <> 'Released'

C: VIEW Not_Released SELECT * FROM movies WHERE release_status <> 'Released'

D: NEW VIEW Not_Released SELECT * FROM movies WHERE release_status <> 'Released'

Q5:
How would you select only the title, release date and release status columns from the view you created in the previous question? (1 mark) [1]

A: Select title, release_date, release_status From VIEW Not_Released

√ B: Select title, release_date, release_status From Not_Released

C: Select title, release_date, and release_status From View Not_Released

D: Select title, release_date, release_status From Not_Released VIEW

Q6: How many movies exist that no longer use their original titles? (2 marks) [2]

A: 187

√ B: 258

C: 24

D: 74

Q7:
What is the most popular movie that was made after 01/01/2000 with a budget of more than $100 000 000? (Hint: Use the popularity field in the Movies table. Larger numbers are more popular.) (2 marks) [2]

A: Pirates of the Caribbean: The Curse of the Black Pearl

√ B: Interstellar

C: Avatar

D: The Dark Knight

Q8: How many movies do not have English as their original language? (2 marks) [2]

A: 387

√ B: 298

C: 315

D: 492

Q9: How many movies in the database were produced by Pixar Animation Studios? (3 marks) [1]

A: 18

√ B: 16

C: 20

D: 14

Q10: How many movies are in the database that are both a Romance and a Comedy? (4 marks) [4]

A: 262

√ B: 484

C: 373

D: 595

Q11:
What is the most popular action movie that has some German in it? (Hint: The German word for German is Deutsch) (3 marks) [3]

A: The Bourne Identity

B: Quantum of Solace

√ C: Captain America: Civil War

D: Mission: Impossible - Rogue Nation

Q12: In how many movies did Tom Cruise portray the character Ethan Hunt? (3 marks) [3]

√ A: 5

B: 4

C: 1

D: 6

Q13: How many times was the actress Cate Blanchett nominated for an Oscar? (3 marks) [3]

√ A: 7

B: 5

C: 2

D: 4

Q14:
How many movies has managed to win Best Picture at the Oscars even though they had a budget of less than $10 000 000? (Hint: The winner is given by a 1 in the “winner” field.) (3 marks) [3]

A: 12

B: 18

√ C: 15

D: 16

Q15:
How many movies contains at least one of the official South African Languages, Afrikaans or Zulu? (3 marks) [3]

A: 12

B: 10

√ C: 8

D: 15

Q16:
What would be the code to change the name of the language with the ‘zh’ iso code in the “language” table to ‘Chinese’? (1 mark) [1]

A: UPDATE languages (language_name = ‘Chinese’) WHERE iso_639_1 = ‘zh’

√ B: UPDATE languages SET language_name = ‘Chinese’ WHERE iso_639_1 = ‘zh’

C: ALTER languages SET language_name = ‘Chinese’ WHERE iso_639_1 = ‘zh’

D: MODIFY languages SET language_name = ‘Chinese’ WHERE iso_639_1 = ‘zh’

Q17: What would be the code to insert a new genre called ‘Sport’ with an id of 10? (1 mark) [1]

A: INSERT INTO genres (genre_id = 10, genre_name = ‘Sport’)

B: INSERT (genre_id, genre_name) INTO genres SET VALUE (10, 'Sport')

C: INSERT genres (genre_id, genre_name) Values (10, 'Sport')

√ D: INSERT INTO genres (genre_id, genre_name) Values (10, 'Sport')

Q18:
You have just watched The Flintstones movie and did not find it very funny. What code would delete the entry that links The Flintstones to the Comedy genre? (2 marks) [1]

√ A: DELETE FROM genremap WHERE genre_id = 35 and movie_id = 888

B: REMOVE ENTRY FROM genremap WHERE genre_id = 35 and movie_id = 888

C: DELETE FROM genremap (genre_id = 35, movie_id = 888)

D: DELETE ENTRY FROM genremap WHERE genre_id = 35 and movie_id = 888

Q19: What code will give me the 10 most recently released movies in the database? (1 mark) [1]

A: SELECT TOP(10) * FROM movies ORDER BY release_date ASC

√ B: SELECT TOP(10) * FROM movies ORDER BY release_date DESC

C: SELECT * TOP(10) FROM movies ORDER BY release_date DESC

D: SELECT * FROM movies ORDER BY release_date ASC LIMIT 10

Q20:
What code would you use to add a column to the language table that could be used for the English names of the different languages? (1 mark) [1]

√ A: ALTER TABLE languages ADD language_english_name varchar(50)

B: UPDATE TABLE languages ADD language_english_name

C: UPDATE TABLE languages APPEND language_english_name varchar(50)

D: ALTER TABLE languages ADD language_english_name

Q21:
Analyse Section (15 Questions, 25 Marks) The practical questions of this section should be answered by analysing the football_players.csv dataset – the recommended tool is Python (Pandas library). In numpy arrays: all data must be of the same type (1 mark) [1]

√ A: True

B: False

Q22: Which of the following statements about numpy arrays is false? (1 mark) [1]

A: It can be modified (changed)

B: It can have more than 2 dimensions

C: It is indexed by a tuple of non-negative integers

√ D: It is exactly the same as lists of lists

Q23: Which of the following statements about pandas dataframes is false? (1 mark) [1]

A: It can be modified (changed)

B: The data type must be the same within a column

C: Data types can differ between columns

√ D: It usually consists of more than 2 dimensions

Q24: How would you select the top left element in a 2-D numpy array a? (1 mark) [1]

A: It is not possible

B: a[:,1]

√ C: a[0,0]

D: a[1,1]

Q25:
How would you add a row of entries 92, 88 and 78 to a 2-D numpy array a with dimensions of 3x3? (1 mark) [1]

A: np.append(a, [[92, 88, 78]], axis=1)

B: np.append(a, 92, 88, 78, axis=0)

C: np.append(a, [92, 88, 78], axis=0)

√ D: np.append(a, [[92, 88, 78]], axis=0)

Q26: How would you select the 5th row of a pandas dataframe, df, with names as indices? (1 mark) [1]

A: df[5]

√ B: df.iloc[4]

C: df.loc[4]

D: df[4]

Q27: How would you create a new column to transform a column 'Age' from years to decades? (1 mark) [1]

√ A: df['decades'] = df['Age'].apply(lambda x: x / 10)

B: df['decades'] = df.applymap(lambda x: x / 10)

C: decades = df['Age'].apply(lambda x: x / 10)

D: df['decades'] = df['Age'].map(def lambda x: x / 10)

Q28: Which Algeria player had the highest overall rating? (2 marks) [2]

√ A: R. Mahrez

B: Y. Brahimi

C: I. Messaoud

D: M. Boulahia

Q29: Which back had the highest rating for 'Sliding tackle'? (2 marks) [2]

A: J. Boateng

√ B: Sergio Ramos

C: M. Hummels

D: R. Woodcock

Q30: Which preferred position type of England has on average the highest overall rating? (2 marks) [2]

A: Midfielder

B: Goalkeeper

C: Forward

√ D: Back

Q31: Brazil's forwards have a higher average overall rating than the backs (1 mark) [1]

A: True

√ B: False

Q32: Which country has the oldest player? (2 marks) [1]

A: Egypt

B: South Africa

√ C: England

D: Mexico

Q33: Which of the following attributes is on average the lowest for goalkeepers? (3 marks) [3]

A: GK reflexes

√ B: GK kicking

C: GK handling

D: GK diving

Q34: Which preferred positions type has the most entries in this dataset? (3 marks) [3]

√ A: Midfielder

B: Forward

C: Back

D: Goalkeeper

Q35: Which player from Portugal, that is younger than 25, has the highest overall rating? (3 marks) [3]

A: Andre Gomes

B: Dany Mota

X C: Joao Mario

D: Bernardo Silva

Q36:
Explain Section (15 Questions, 25 Marks) The practical questions of this section should be answered by using the IPL Power BI file. Note that some measures (Batsman Runs, Batsman Balls, Batsman Strike Rate, Batsman Average) have already been calculated – please use them to answer the questions. Also all questions relate to Royal Challengers Bangalore Batsman, therefore a page filter of batting_team = "Royal Challengers Bangalore" has already been added. Which one of the following data sources can Power BI NOT connect to? (1 mark) [1]

A: SQL Database

B: CSV file

C: JSON file

√ D: Docx file

Q37: Which one of the following statements about DAX is incorrect? (1 mark) [1]

A: It can be used to create calculated columns and measures

X B: It cannot modify or insert data

C: It works on column values

D: It can be used to calculate rows

Q38: What does the following DAX measure do: (1 mark)

CALCULATE(COUNT(table1[col1]), table[col1] = "x")

[1]

A: Multiples the number of rows in col1 with "x"

B: Adds the rows in col1 to "x"

√ C: Counts the number of rows in col1 that is equal to "x"

D: Calculates the number of times col1 is not equal to "x"

Q39:
Which type of visual in Power BI would be best suited if we want to display data per category, when we have many categories with long names? (1 mark) [1]

A: Pie Chart

B: Line Chart

√ C: Bar Chart

D: Column Chart

Q40:
Which of the following 3 actions can be performed in the formatting section of a visual in Power BI? (1 mark)

1. Rounding data
2. Changing the data type
3. Changing font size

[1]

√ A: Only 1 and 3

B: All of them

C: Only 1

D: Only 2 and 3

Q41: Which one of the following is not regarded as best practice when building a dashboard? (1 mark) [1]

A: Don't clutter your dashboard

B: Group data logically

C: Making it relevant to the audience

√ D: At least 5 different type of visuals per page

Q42: Who scored the most runs for RCB in 2016? (hint: use the Batsman Runs measure) (2 marks) [2]

A: AB de Villiers

X B: CH Gayle

C: Virat Kohli

D: SR Watson

Q43:
How many runs did JH Kallis score for RCB in 2010? (hint: use the Batsman Runs measure) (2 marks) [2]

A: 361

√ B: 572

C: 199

D: 0

Q44:
Who has the 4th highest batting average for RCB against Kings XI Punjab (from only the players that scored more than 100 runs against them)? (hint: use the Batsman Average measure) (2 marks) [2]

√ A: R Dravid

B: JH Kallis

C: V Kohli

D: CH Gayle

Q45:
In which year did V Kohli score his most runs for RCB? (hint: use the Batsman Runs measure) (2 marks) [2]

A: 2014

√ B: 2016

C: 2017

D: 2015

Q46:
Against which team did JA Morkel have the lowest strike rate when batting for RCB? (hint: use the Batsman Strike Rate measure) (2 marks) [2]

A: Kings XI Punjub

B: Kolkata Knight Riders

√ C: Rajasthan Royals

D: Chennai Super Kings

Q47:
Who scored the 5th most runs for RCB after 22 April 2015? (hint: use the Batsman Runs measure) (2 marks) [2]

A: TM Head

B: KM Jadhav

C: SN Khan

√ D: Mandeep Singh

Q48:
From only the players that scored more than 100 runs, who had the highest batting index for RCB, where batting index is defined as Strike Rate plus Average? (hint: use the Batsman Average and Batsman Strike Rate measures) (2 marks) [2]

A: V Kohli

B: R Dravid

C: CH Gayle

√ D: AB de Villiers

Q49: How many sixes did RCB hit in 2011? (2 marks) [2]

√ A: 94

B: 129

C: 105

D: 118

Q50: What proportion of runs has CH Gayle scored in sixes over his IPL career for RCB? (3 marks) [3]

A: 55%

B: 50%

C: 40%

√ D: 45%

Student: Damian Vather

Programme: Data Science

Test: EDSA Final Exam - Part 2 :

Mark: 74.74 %

Notes to Facilitators

- Clear feedback will be given verbally to the Assessor/Moderator regarding the marking system.
- Underlined Text indicates delegate's selection. If there is no Underlined Text, delegate did not select an option.
- Red text indicates the expected answer
- Red √ where delegate selected expected options
- Red X where delegate selected incorrect options

Feedback for delegate

Not available

Q1:
NLP (10 Questions, 22 Marks) The practical questions of the next 12 questions should be answered using the Essay_data.csv file. This csv file contains a personality profile, together with an essay written by an individual with that specific personality type. Once you have imported the data frame, use the dropna() function to remove rows containing missing values. How many bi-grams can be created from the following sentence, after stopwords have been removed and punctuation has been replaced by a single white space: (2 marks)
I'm a part-time student @explore-software. [2]

X A: 6

B: 5

C: 7

D: 8

Q2:
Which ratio best relates the number of intuitive (N) students to the number of sensing (S) students? (2 marks) [2]

A: 4:6

B: 6:4

√ C: 7:3

D: 3:7

Q3:
Remove all punctuation from the essays and convert it to lower case. What is the 10th character in the first essay? (2 marks) [2]

√ A: '4'

B: ' '

C: 't'

D: 'filled'

Q4: Tokenize the essays. How many tokens are in the 17th essay? (2 marks). [2]

A: 128

B: 440

C: 94

√ D: 337

Q5: How does lemmatization differ from stemming? (1 mark) [1]

A: They operate the same way – they just come from different libraries (scikit-learn and NLTK).

√ B: Lemmatization always results in actual words, while stemming can result in non-existent words

C: Stemming always results in actual words, while lemmatization can result in non-existent words.

D: Lemmatization cuts off part of the word, while stemming considers the morphological analysis of the words.

Q6: Use the *SnowballStemmer* to stem the word 'experiences'. What is the output? (1 mark) [1]

A: 'experienc'

B: 'experience'

C: 'experiences'

√ D: 'experi'

Q7: Remove all the stop words. What is the 24th token in the 81st essay? (2 marks) [2]

√ A: 'times'

B: 'evidently'

C: 'working'

D: 'selfconfidence'

Q8: How many unique words are in these essays (after we have removed the stopwords)? (2 marks) [1]

A: 3494

B: 3812

C: 4087

√ D: 3406

Q9:
Create a bag of words and use it to determine how many times ‘time’ was mentioned in the 56th essay? (2 marks) [2]

A: 2

B: 1

√ C: 3

D: 0

Q10:
Words that appear at least twice, account for what percentage of the total number of words in the essays (2 marks) [2]

A: 48%

B: 81%

√ C: 90%

D: 62%

Q11: What is the most commonly mentioned word by ENFJ personalities? (2 marks) [2]

A: ‘data’

B: ‘group’

√ C: ‘team’

D: ‘academy’

Q12:
Create a new column in the data frame containing the bi-grams from each essay. What is the 109th bi-gram in the 70th essay? (2 marks) [2]

A: (‘quite’, ‘well’)

B: (‘work’, ‘quite’)

C: (‘better’, ‘certain’)

X D: (‘may’, ‘better’)

Q13:
Classification Section (22 Questions, 40 Marks) The next 11 questions are based on the medical claims dataset (claims_data.csv): What proportion of individuals in this dataset would be classified as overweight or obese (BMI of greater than 25)? (2 marks) [2]

A: 85%

√ B: 82%

C: 15%

D: 73%

E: 18%

Q14:
Is the Poisson distribution a good choice to model the distribution of the number of children in this dataset? (1 mark) [1]

A: No, because the Poisson only applies to positive integers, so cannot accommodate observations with 0 children.

√ B: No, the variance is significantly higher than the mean, suggesting overdispersion relative to the Poisson distribution.

C: No, the variance is significantly lower than the mean, suggesting underdispersion relative to the Poisson distribution.

D: No, the Poisson is inappropriate, as it is a continuous distribution while the number of children is a discrete variable.

E: Yes, the Poisson is a good choice for count data such as the number of children in a family.

Q15:
If we assumed that age of this group was normally distributed, then given the mean and standard deviation of age in the data set, calculate the number of individuals we would expect to be aged 60 or older (use 60 exactly as the cutoff point on the distribution, and round to the nearest integer). Then compare this with the number actually aged 60 or older. Which of the following is true? (3 marks) [3]

A: There are 21 fewer individuals 60 or older than the normal distribution would suggest.

B: The two are exactly equal!

C: There are 7 fewer individuals 60 or older than the normal distribution would suggest.

√ D: There are 21 more individuals 60 or older than the normal distribution would suggest.

E: There are 7 more individuals 60 or older than the normal distribution would suggest.

Q16: Create a joint plot on the age and BMI variables. What summarises best what you see? (2 mark) [2]

A: There is not an easily discernible pattern in the plot, but the correlation coefficient is 0.11 which is statistically significantly different from zero, suggesting that older people tend to have higher BMIs.

B: There is not an easily discernible pattern in the plot, but the correlation coefficient is 0.11 which is statistically significantly different from zero, suggesting that older people tend to have lower BMIs.

C: There is not an easily discernible pattern in the plot, and the correlation coefficient is 0.11 which is not statistically significantly different from zero.

X D: There is a clearly discernible pattern in the plot, with a tight clustering and upward trend showing that older people tend to have higher BMIs, confirmed by a correlation coefficient of 0.11 which is statistically significantly different from zero.

E: There is a clearly discernible pattern in the plot, with a tight clustering and downward trend showing that older people tend to have lower BMIs, confirmed by a correlation coefficient of 0.11 which is statistically significantly different from zero.

Q17:
Use the appropriate model from the sklearn library (with default parameters unless specified otherwise) to fit a logistic regression model to the data, with `insurance_claim` as your target variable, using all other fields apart from `claim_amount` and creating dummy variables for the categorical variables in the data, dropping the first in each instance. Do a test-train split holding out 33% of the data for the test set, using a random seed of 42 for the split. Convert your target variable to a binary 0 or 1, where 1 indicates that there was a claim. What proportion of claim indicators in the test set are correctly predicted? (3 marks) [3]

A: 81%

X B: 85%

C: 91%

D: 87%

E: 86%

Q18:
Now fit another logistic regression, this time using the statsmodels library to do so, with default parameters. Be sure to add a constant to your X matrices, both train and test (you might want to check the statsmodels documentation for the add_constant function). Which of the following best summarises the results? (3 marks) [3]

X A: Apart from age, BMI and smoker status, none of the other features are statistically significant predictors of an insurance claim.

B: All features seem to be significant predictors of the likelihood of a claim.

C: Age, BMI, number of children and smoker status significantly affect the likelihood of an insurance claim. Number of steps and sex are not significant. There appear to be some regional effects but these are not strongly significant.

D: Age, sex, BMI and smoker status significantly affect the likelihood of an insurance claim. Number of steps and number of children are not significant. There appear to be some regional effects but these are not strongly significant.

E: Age, sex, BMI, number of children and smoker status significantly affect the likelihood of an insurance claim. Number of steps is not significant. There appear to be some regional effects but these are not strongly significant.

Q19: What is the primary reason for using random forests instead of a single decision tree? (1 mark) [1]

A: Decision trees suffer from high variance, and random forests reduce this variance by averaging multiple trees each fitted to a subset of the observations and ensuring these trees are decorrelated by using only a subset of the available predictors.

B: Decision trees suffer from high bias, and random forests reduce this, but at the expense of higher variance.

C: There is no good reason in principle to prefer a random forest over a single decision tree; it depends on the data.

X D: Decision trees suffer from high variance, and random forests reduce this variance by averaging multiple trees each fitted to a subset of the observations.

E: There is much more firewood in a whole forest than in a single tree.

Q20:
Now fit a random forest with 100 trees and a random seed of 101, and default parameters for the rest. Which of the following sets out the number of false negatives and false positives in the confusion matrix on the test data? (2 marks) [2]

A: FN = 4, FP = 16

B: FN = 245, FP = 177

C: FN = 12, FP = 12

√ D: FN = 16, FP = 4

E: FN = 177, FP = 245

Q21:
Fit Support Vector Machine models to the training data, using respectively the radial, sigmoid and linear kernels with default parameters. Which model yields the best accuracy on test data? (2 marks) [2]

A: Linear

B: Radial and linear yield very similar accuracies

C: Sigmoid

D: Sigmoid and linear yield very similar accuracies

X E: Radial

Q22: With respect to a SVM, which of the following is true? (1 mark) [1]

A: The penalty parameter has no influence on the accuracy of the model on training data, only on test data.

B: Training accuracy can be improved by decreasing the value of the penalty parameter.

C: The penalty parameter cannot be varied using sklearn.

√ D: Training accuracy can be improved by increasing the value of the penalty parameter.

E: The default value of the penalty parameter is optimal; we can't improve the model fit on training data by either increasing or decreasing it.

Q23:
The next 4 questions are based on the IPL match data (matches.xlsx). The indicator dl_applied refers to weather-shortened matches in which the Duckworth-Lewis method was applied to determine the winner. In what proportion of matches did this happen? (2 mark) [2]

A: 50%

B: 5%

C: 25%

√ D: 2.5%

E: 0.25%

Q24: What proportion of matches was won by the team who batted first? (2 mark) [2]

A: 48%

√ B: 45%

C: 52%

D: Impossible to determine from the data

E: 55%

Q25:
We define a close match as one which was won by 20 runs or less, or by 4 wickets or less. We want to build a model to predict whether or not a game will be close based on the following three features:

- whether the match was played in the month of April, or not
- whether the toss winners chose to bat or field
- whether or not the Duckworth-Lewis method was applied

Create these features. Which of the following tuples correctly enumerates respectively the number of April games and choices to field first across the data set? (3 marks) [3]

A: (297, 273)

B: None of the above

C: (339, 363)

D: (339, 273)

√ E: (297, 363)

Q26:
Build a decision tree classifier on these features, using a train-test split with a 75:25 weight and a random seed of 999. Which of the following is the most accurate reflection of the confusion matrix on the test data? (3 marks) [3]

A: The model predicts that the vast majority of games will be close wins, hence we have few false negatives but many false positives.

B: The model predicts that the vast majority of games will not be close wins, hence we have few false negatives but many false positives.

C: The model predicts that the vast majority of games will be close wins, hence we have few false positives but many false negatives.

D: The model predicts that the vast majority of games will be close wins, hence we have few false negatives but many false positives.

√ E: The model predicts that the vast majority of games will not be close wins, hence we have few false positives but many false negatives

Q27:
The next 3 questions are based on the FIFA players dataset (football_players.csv). Please note that you may have some difficulties importing this file into pandas, and you may need to do some research to figure out how to do so successfully. What is the most common Overall score for players in the database? (1 mark) [1]

A: 67

√ B: 66

C: 68

D: 93

E: 64

Q28:
Construct a dataset which is a subset of players who can play in central defence (i.e. who have 'CB' somewhere in their `Preferred Positions` field). Split this group into three:

1. World Class: overall score of 80 or more
2. Good: overall score of 70-79
3. Mediocre: overall score below 70

Now build a random forest classifier with default parameters apart from setting to 500 trees and setting the random seed to 1971, on ALL of the data for these central defenders, where the target variable is the classification into one of the three classes defined above, and the candidate features are all other numerical variables. In descending order, which are the five most important features that emerge from this model? (Hint: search sklearn random forest documentation for feature importance if you don't have any idea how to establish this.) (3 marks) [3]

A: Composure, Sliding Tackle, Aggression, Short passing, Marking

B: Marking, Reactions, Composure, Standing Tackle, Sliding Tackle

√ C: Standing tackle, Marking, Interceptions, Sliding Tackle, Reactions

D: Standing tackle, Sliding Tackle, Composure, Marking, Heading accuracy

E: Composure, Standing Tackle, Marking, Heading accuracy, Sliding Tackle

Q29:
Why do we generally not use all the data to fit models as we did in the previous question, but rather perform a train-test split or cross-validation? (1 mark) [1]

A: Because Dewald gets mad if we don't.

B: Although it's considered best practice to split the data into test and training sets, this is not actually required in many circumstances.

C: To simultaneously minimise bias and variance.

D: To avoid underfitting to the data, and hence improve our chances of generalising to unseen data.

√ E: To avoid overfitting to the data, and hence improve our chances of generalising to unseen data.

Q30:
Split the data into test and training sets, with 33% of the data reserved for the test set and a random seed of 911. Compare k nearest neighbours (KNN) models with k varying from 1 to 5. Which k gives rise to the best F1 score for the world class good groups respectively? (3 marks) [3]

A: World class: k=1, Good: k=4

B: World class: k=4, Good: k=5

√ C: World class: k=5, Good: k=4

D: World class: k=1, Good: k=2

E: World class: k=5, Good: k=5

Q31:
Split the data into test and training sets, with 33% of the data reserved for the test set and a random seed of 911. Compare k nearest neighbours (KNN) models with k varying from 1 to 5. Which k gives rise to the best F1 score for the world class good groups respectively? (3 marks) [3]

X A: World class: k=4, Good: k=5

B: World class: k=5, Good: k=5

C: World class: k=5, Good: k=4

D: World class: k=1, Good: k=4

E: World class: k=1, Good: k=2

Q32: Which of the following is an accurate description of logistic regression? (1 mark) [1]

A: A coefficient of below 1 in a logistic regression implies a negative contribution to the probability of being in the target class.

√ B: Logistic regression fits a linear model to the log odds ratio, which is the log of the probability of being in a class as a proportion of the probability of not being in that class.

C: Logistic regression fits a linear model to the log odds ratio, which is the probability of being in a class as a proportion of the probability of not being in that class.

D: Logistic regression is not a linear model, unlike linear regression.

E: Logistic regression fits a linear model to the odds ratio, which is the probability of being in a class as a proportion of the probability of not being in that class.

Q33:
Which of the following are true of the k-nearest neighbours (KNN) algorithm applied to an n-dimensional feature space? (1 mark)

I. For a new test observation, the algorithm looks at the k training observations closest to it in n-dimensional space and assigns it to the majority class among those k observations.
II. For a new test observation, the algorithm looks at the k training observations closest to it in n-dimensional space and assigns it proportionally to each class represented in those k observations.
III. KNN models tend to perform poorly in very high dimensions.
IV. KNN models are well-suited to very high-dimensional data.
V. The K in KNN stands for Kepler, the scientist who first proposed the algorithm

[1]

√ A: I and III.

B: I, III and V.

C: I only.

D: II and IV.

E: I, IV and V.

Q34: What is a hyperparameter? (1 mark) [1]

A: A parameter whose value is set before the model-fitting process begins.

B: None of the above.

C: A parameter which can only be set by grid search.

D: A model parameter which has more than one dimension.

X E: Machine learning terminology for a model parameter.

Q35:
Regression Section (20 Questions, 30 Marks) For this section please follow the steps below before attempting the questions. Load the data set titled 'rand-dollar.csv' as follow:

pd.read_csv('rand-dollar.csv', index_col=0)

Separate the data set into X (features) and y (targets). Our target variable will be ZAR/USD, with all other variables being the predictors. Create an 80/20 split between train and test sets. The training data should be the first 80% of the data, with the final 20% being used in the test set:

train_test_split(X, y, test_size=0.2, shuffle=False)

Train a simple linear regression model to predict the 'ZAR/USD' using only 'Value of Exports (ZAR)' as the predictor variable:

```
from sklearn.linear_model import LinearRegression
lm = LinearRegression( )
lm.fit(...)
```

What is the value of the intercept of the model? (2 marks) [2]

A: 3.99

B: 8.67e-5

C: -1.24

√ D: 3.29

Q36: How do we interpret the intercept? (1 mark) [1]

√ A: The value of ZAR/USD will be equal to this value when exports are zero

B: The value of ZAR/USD will be equal to zero when exports are equal to this value

C: This is the average value of ZAR/USD

D: The maximum value of exports is equal to this value

Q37: What is the value of the slope of this model? (2 marks) [2]

A: 86 800 000

√ B: 8.68 e -5

C: 3.29

D: 8.68

Q38: How do we interpret the slope of the model? (1 mark) [1]

A: A decrease of this many units in exports results in an increase of 1 unit in ZAR/USD

B: An increase of 1 unit in exports results in an increase of this many units in ZAR/USD

X C: An increase this many units in exports results in an increase of 1 unit in ZAR/USD

D: A decrease of 1 unit in ZAR/USD results in an increase of this many units in exports

Q39:
What is the predicted value of the exchange rate in a month where exports total R100 000? (2 marks) [2]

A: R90.07 / $1.00

√ B: R11.97 / $1.00

C: R4.16 / $1.00

D: $11.97 / R1.00

Q40: What is the MSE of the model on the test set? (2 marks) [2]

A: 1.89

B: -8.45

C: 6

√ D: 8.22

Q41: What is the R-squared value of the model on the test set? (2 marks) [2]

A: 9

√ B: -8.45

C: 8.22

D: -4.25

Q42: What would a negative R-squared test value imply? (1 mark) [1]

A: The data needed to be standardized before the model was trained

B: R-squared has no implication when using a test set

C: The model is worse at predicting the target variable than if a constant line y = intercept was used

X D: The relationship between predictor and response is negative

Q43: What is the predicted value for August 2017? (2 marks) [2]

√ A: R12.25 / $1.00

B: $12.25 / R1.00

C: R16.51 / $1.00

D: R7.13 / $1.00

Q44: What is the absolute error for this prediction? (1 mark) [1]

√ A: R0.98

B: R2.42

C: R0.49

D: R1.98

Q45:
Now revert back to using the full original data set. Use the df.corr() function to find the correlations between the predictors and the target variable.

Which variable has the weakest linear relationship with the ZAR/USD exchange rate? (2 marks) [2]

X A: Claims on Non-residents

B: Savings Rate

C: Lending Rate

D: Consumer Price Index

Q46: Which variable has the strongest linear relationship with the ZAR/USD exchange rate? (2 marks) [2]

√ A: Consumer Price Index

B: IMF Reserve Position (USD)

C: Savings Rate

D: Claims on Non-residents

Q47:
Before answering the following questions, make sure to perform the steps outlined below:

- Split the original dataframe into X (features) and y (targets)
- Standardize the entire X matrix
- Create X_train, X_test, y_train, y_test using the same chronological 80/20 split as before
- Train two models “ridge” and “lasso” which use ridge regression and LASSO respectively (in the case of the lasso model, set alpha=0.01, use default parameters for the ridge model) (2 marks)

What is the training MSE of the Ridge model? (2 marks) [2]

A: 0.632

B: 0.047

C: 0.579

√ D: 0.040

Q48: What is the training MSE of the LASSO model? (2 marks) [2]

A: 0.040

√ B: 0.047

C: 0.579

D: 0.632

Q49: What is the testing MSE of the Ridge model? (1 mark) [1]

A: 0.047

B: 0.579

√ C: 0.632

D: 0.640

Q50: What is the testing MSE of the LASSO model? (1 mark) [1]

A: 0.047

B: 0.632

√ C: 0.579

D: 0.640

Q51:
Based on the values of the Ridge model's variable coefficients, which indicator is the best predictor of the target variable? (1 mark) [1]

A: Value of Imports (ZAR)

B: Government Bonds

C: Liabilities to Non-residents (USD)

√ D: Value of Exports (ZAR)

Q52:
Based on the values of the Ridge model's variable coefficients, which indicator is the worst predictor of the target variable? (1 mark) [1]

A: Value of Exports (ZAR)

B: Liabilities to Non-residents (USD)

X C: Government Bonds

D: Value of Imports (ZAR)

Q53:
Based on the values of the LASSO model's variable coefficients, which indicator is the best predictor of the target variable? (1 mark) [1]

A: Government Bonds

√ B: Value of Imports (ZAR)

C: Value of Exports (ZAR)

D: Liabilities to Non-residents (USD)

Q54: How many variables have coefficients equal to zero in the LASSO model? (1 mark) [1]

√ A: 4

B: 3

C: 9

D: 0