REPUBLIC OF ALBANIA

UNIVERSITY OF TIRANA

FACULTY OF NATURAL SCIENCES

Mars Craters

Data Manipulation and Visualization

Master of Science in Informatics

Data Mining

Content:

List of tables:

List of figures:

1. What is Data Analysis?

Data Analysis is the process of collecting, modeling and analyzing data to extract knowledge that supports decision making. There are several methods and techniques to perform analysis depending on the industry and the purpose of the analysis. All these different methods for data analysis are mainly based on two main areas: quantitative methods and qualitative research methods.

The purpose of Data Analysis is to extract useful information from the data and make a decision based on the analysis of this data. A simple example of data analysis is every time we make a decision in our daily life is thinking about what happened last time or what will happen by choosing that specific decision we want to make.

2. Steps in Data Analysis:

With the right data analysis tools and process, what was once voluminous and incomprehensible information becomes a clear and simple decision. To improve data analysis skills and simplify decisions, we take these five steps in the data analysis process:

Defining research questions:

1. Setting clear measurement priorities
2. Data collection
3. Data analysis
4. Interpretation of results

Description of steps:

1. In analyzing organizational or business data, we need to start with the right questions. Questions should be measurable, clear and concise. It is important to design qualified questions or disqualify possible solutions to the problem.

2. The second step consists of two sub-activities that need to be defined:

   a) Decide what to measure.
   b) Decide how to measure it.

3. With the question clearly defined and priorities set, now is the time to gather data.

4. Once we have gathered the right data to answer the question from Step 1, it is time for deeper data analysis. We start by manipulating the data in a number of different ways and finding correlations, the relationships between them.
5. After analyzing the data and perhaps conducting further research, it is finally time to interpret the results that will lead to the decision making and answering the question.

3. Course: Data Management and Visualization

Description:

Whether used to personalize ads for millions of website visitors or to simplify ordering in a small restaurant, data is becoming more integral Too often, we are not sure how to use data to find answers to questions that will make us more successful in what we do. In this course, we will discover what data is and think about what questions we can answer from that data. Based on the existing data, we will learn to develop a research question, describe the variables and the relationships between them, calculate the basic statistics and clearly present our results. At the end of the course, we will be able to use powerful data analysis tools - SAS / Python - to manage and visualize data,

Python is a programming language where we can write our Python code in a programming environment that provides something like this. The selected IDE isPyCharm. Anaconda simplifies the installation process including Python, a programming environment, packages, and other tools in one installation file.

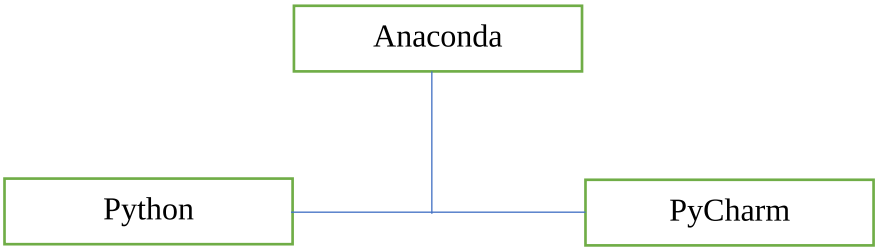


Figure 1: Creating the environment for data mining development

4. Anaconda, Python & PyCharm

Anaconda:

Anaconda is a "distribution" of Python and R programming languages for scientific calculations (data science, machine learning applications, data processing, predictive analytics, etc.), which aims to simplify the management and deployment of packages. The distribution includes data science packages suitable for Windows, Linux and macOS. Package versions in Anaconda are managed by the conda package management system. This package manager was developed as a separate open-source package for Python.

Python:

Python started as an open-source programming language although it was not originally used to perform data analysis. Pandas and other specialized bookstores have begun to change this. Python is always supported and changed periodically and is always evolving with new versions which are updated from time to time for usage purposes. Python is known for its simplicity in the world of programming. This also applies to analytical data.

PyCharm:

PyCharm is an integrated development environment (IDE) used in programming, especially for the Python language. PyCharm Professional Edition helps us analyze our data with Python. Just create a project, add the data and then start analyzing it. Data visualization is an essential step in any data analysis, PyCharm helps us by showing us the plots within the IDE. PyCharm also keeps track of the last created, making it easy to detect changes.

References to install the components listed above:

Anaconda:	https://docs.anaconda.com/anaconda/install/windows/
Python:	https://www.python.org/downloads/
PyCharm:	https://www.jetbrains.com/help/pycharm/installation-guide.html

Configuring PyCharm with Anaconda for data management and manipulation:

https://docs.anaconda.com/anaconda/user-guide/tasks/pycharm/

Metric:

Python version:	3.8.12
PyCharm version:	2021.2.3 Professional Edition
Conda Version:	conda 4.10.1

5. Entry

The multi-crater surface on Mars was created between 4.2 and 3.8 billion years ago during the period of "heavy bombardment" (i.e. the impacts of asteroids, protoplanets, and comets). Craters are an exogenous planetary process that contribute to the superficial change of bodies in the solar system. Craters appear almost all over the surface of Mars, and they serve as information to understand the properties of the Martian crust, as well as the age and time of these events. Outdated, automated and inaccurate datasets of craters data were created on Mars, Dataset is a global database for Mars containing 384,344 complete statistical records for diameters D> 1 km. This detailed database includes location and size,

Information:

The database is from a new global database of craters on Mars, created by Stuart Robbins as part of his doctoral thesis (Doctorate) conducted between 2008 and 2011. It contains individual features and parameters from 384, 344 craters. The global Stuart Mars Crater Database is extensive because it contains a complete set of information for large craters as well as smaller craters less than 1 km in diameter. The global database contains craters with measured crater depth ranging from -0.42 km to 4.95 km. The diameter of the crater is between 1 km and 1164.22 km. Crater latitude and longitude were measured in degrees north (longitude) and east degrees (latitude), respectively. The recorded number of crater layers is from 0 to 5. Craters were mainly classified into different morphological groups; mainly simple and complex morphological categories.

Unique identifier: CRATER_ID (crater id, unique number, not repeated)

Variables:

Naming the variable	Description
CRATER_ID	crater id, based on which area of the planet it is located (1/16)
LATITUDE_CIRCLE_IMAGE	latitude, derived from the center of the square with the vertices selected to identify the edges of the crater (units: degrees)
LONGITUDE_CIRCLE_IMAGE	longitude, derived from the center of the square with the vertices selected to identify the edges of the crater (units: degrees)
DIAM_CIRCLE_IMAGE	crater diameter (unit: km)
DEPTH_RIMFLOOR_TOPOG	the average height of each of the N points defined along (or inside) the edges of the crater (units: km)
MORPHOLOGY_EJECTA_1	morphology classification,
MORPHOLOGY_EJECTA_2	morphology of the layers themselves

MORPHOLOGY_EJECTA_3	the shape of some of the layers
NUMBER_LAYERS	the maximum number of connected layers in each direction that could be reliably identified

Table 1: Description of data variables

Procedure:

From existing satellite imagery and databases, Stuart used modern digital techniques and algorithms to identify craters and determine their basic properties such as depth, diameter, location (latitude and longitude), morphology, and number of layers. The crater identification and classification procedure was performed primarily on ArcGIS software using infrared image observations captured by a THEMIS aboard NASA's Mars Odyssey 2001 spacecraft.

Research questions:

Our main research question is to determine if the depth of the crater is related to the diameter of the crater and to verify if this connection depends on the location of the crater on Mars (the width of the crater). The depth of the crater is the response variable while the diameter of the crater is the explanatory variable. The width of the crater will be used as a moderator.

Crater depth is a quantitative variable that measures the average height from the last point of depth to the highest point of depth (units are in kilometers). Crater diameter is also a quantitative variable which measures the size of craters (units are in km). Location was measured by two quantitative variables - crater width and crater length. For the current analysis, the width of the crater was divided into four categories; each contains an interval of 45 degrees from north to south. This new variable is called MARS_REGION. Crater depth and diameter will be used as quantitative variables during regression analysis.

6. Task 1: Create our first program in Python

We create a new project in PyCharm, in this case named MarsCraters. We configure the project with certain python interpreter provided by anaconda by adding a new conda environment to our project. To do this we are giving a short tutorial link according to the steps:How to configure PyCharm with Anaconda.

In this project we create a directory to store our dataset. We created a directory called the dataset.Mars Craters Datasetis an open-source dataset and can be downloaded. Once we download the dataset we place it in our PyCharm project in the dataset directory we created earlier.

To save the python environment configurations we create a python environment using the command:

```
python -m venv dataset_env
```

After creating the python environment we activate it with the command: activate.

```
python -m venv dataset_env
.\dataset_env\Scripts\activate
```

We create a new python file to write our code in python. We create a python file named run.py. Python is one of the programming languages that offers numerous libraries, including data analysis and machine learning. One of the most popular libraries is pandas and another library is numpy.

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it provides data structures and operations for manipulating numerical tables and time series.

Numpy is a library for the Python programming language, adding support for large, multi-dimensional vectors and matrices, along with a large collection of high-level mathematical functions to operate on these vectors.

A code written in python starts with importing libraries to be used during the code. Among the libraries that are the most used are exactly the libraries mentioned above. The next step is to read the dataset and this is done exactly by a method provided by the pandas library and the method name is read_csv. This method makes it possible to read a database in csv format. The method takes as a parameter the path of a dataset and can take other parameters such as. low_memory.

To verify that the dataset has been read correctly we can try to display information from that dataset, we can display the number of records / rows that the dataset contains, in the same way we can read the number of variables / columns that the dataset contains, which corresponds to the number of variables we can get information from and study. In the same way we can do other operations with these variables, such as. the number and percentage / frequency of a given variable, we can create smaller groups to study within this dataset and other operations.

```
# import packages that we need
import pandas as pd
import numpy as np

# print (pd .__ version__)
# read from dataset
data = pd.read_csv ('dataset / marscrater_pds.csv', low_memory = False)

# number of observations (rows)
print ('Number of records:' + str (len (data)))
# number of variables (columns)
print ('Number of variables:' + str (len (data.columns)))

# checking the variable data type
print ('Crater ID Variable Type:' + str (data ['DIAM_CIRCLE_IMAGE'].
dtype))

# setting variables to numeric
# data ['DIAM_CIRCLE_IMAGE'] = pd.to_numeric (data
['DIAM_CIRCLE_IMAGE'])
# print ('Diameter Variable Type:' + str (data ['DIAM_CIRCLE_IMAGE'].
dtype))

# counts and percentages (ie frequency distributions) for this variable
count_1 = data ['DIAM_CIRCLE_IMAGE']. value_counts (sort = False)
print (count_1)

percentage_1 = data ['DIAM_CIRCLE_IMAGE']. value_counts (sort = False,
normalize = True)
print (percentage_1)

frequency_1 = data.groupby ('DIAM_CIRCLE_IMAGE'). size ()
print (frequency_1)

# subset data to craters with diameter between 50 and 80 kms
subset_1 = data [(data ['DIAM_CIRCLE_IMAGE']> = 50) & (data
['DIAM_CIRCLE_IMAGE'] <= 80)]

subset_2 = subset_1.copy ()

# frequency distributions on new sub2 data frame
print ('Counts for Diameter Circle:')
count_2 = subset_2 ['DIAM_CIRCLE_IMAGE']. value_counts (sort = False)
print (count_2)

# upper-case all DataFrame column names - place afer code for loading
data aboave
data.columns = list (map (str.upper, data.columns))

# bug fix for display formats to avoid run time errors - put after code
for loading data above
pd.set_option ('display.float_format', lambda x: '% f'% x)
```

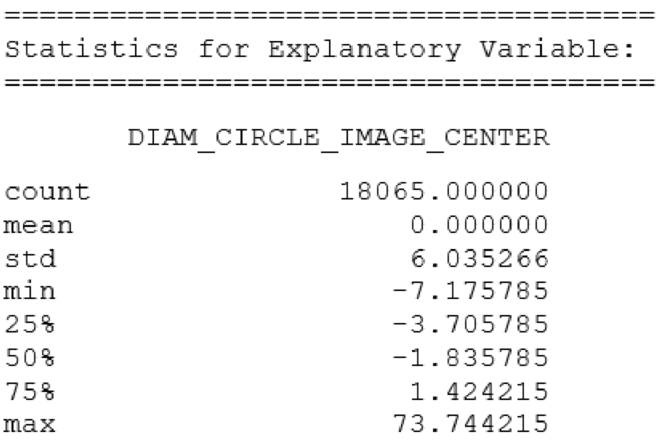
As part of the task we will try to find and construct a basic linear regression model for the relationship between crater diameter (explanatory variable) and crater depth (response variable). Explanatory and response variables are both quantitative.

Below are the python code used in the regression analysis and a brief report to summarize the results of the regression model.

```
data ['DIAM_CIRCLE_IMAGE'] = pd.to_numeric (data ['DIAM_CIRCLE_IMAGE'],
errors='coerce') data ['DEPTH_RIMFLOOR_TOPOG'] = pd.to_numeric (data
['DEPTH_RIMFLOOR_TOPOG'], errors='coerce')

working_subset_1 = data [(data ['NUMBER_LAYERS']> 0) & (data
['DIAM_CIRCLE_IMAGE']> 0) & (data ['DIAM_CIRCLE_IMAGE'] <= 100) & (data
['DEPTH_RIMFLOOR_TOPOG']> 0) & (data ['DEPTH_RIMFLOOR_TOPOG'] <= 3)]
working_subset_2 = working_subset_1.copy () working_subset_2
['DIAM_CIRCLE_IMAGE_CENTER'] = (working_subset_2 ['DIAM_CIRCLE_IMAGE']
- working_subset_2 ['DIAM_CIRCLE_IMAGE'] .mean ())
```

```
# explanatory variable
print('===================================')
print('Statistics for Explanatory Variable:')
print('===================================')
print(working_subset_2 [['DIAM_CIRCLE_IMAGE_CENTER']]. describe ())

# regression model
print('============================================= ========= ')
print('Regression Model between crater DIAMETER and crater DEPTH:')
print('============================================= ========= ')
regresion_1 = stats_formula.ols ('DEPTH_RIMFLOOR_TOPOG ~
DIAM_CIRCLE_IMAGE_CENTER', date= working_subset_2) .fit ()
print(regresion_1.summary ())

slope, intercept, r_value, p_value, sdt_err = stats.linregress
(working_subset_2 ['DIAM_CIRCLE_IMAGE_CENTER'], working_subset_2
['DEPTH_RIMFLOOR_TOPOG'])

graph_1 = seaborn.lmplot (x='DIAM_CIRCLE_IMAGE_CENTER',
y='DEPTH_RIMFLOOR_TOPOG', height=9, fit_reg=True, line_kws=
{'color':'red'}, date= working_subset_2)

pyplot.xlabel ('CRATER DIAMETER CENTER (km)')
pyplot.ylabel ('CRATER DEPTH (km)')
pyplot.title ('Relationship between crater diameter and crater depth',
fontsize=20, fontweight='bold') graph_1.set (xlim= (-20, 80))
graph_1.set (ylim= (0, 4))
pyplot.xticks (np.arange (-20, 80+200, 20))
pyplot.yticks (np.arange (0, 4+1, 1))
pyplot.tick_params (axis='both', labelsize=19)
pyplot.show () graph_1.savefig ('output_regresion_model.png')
```

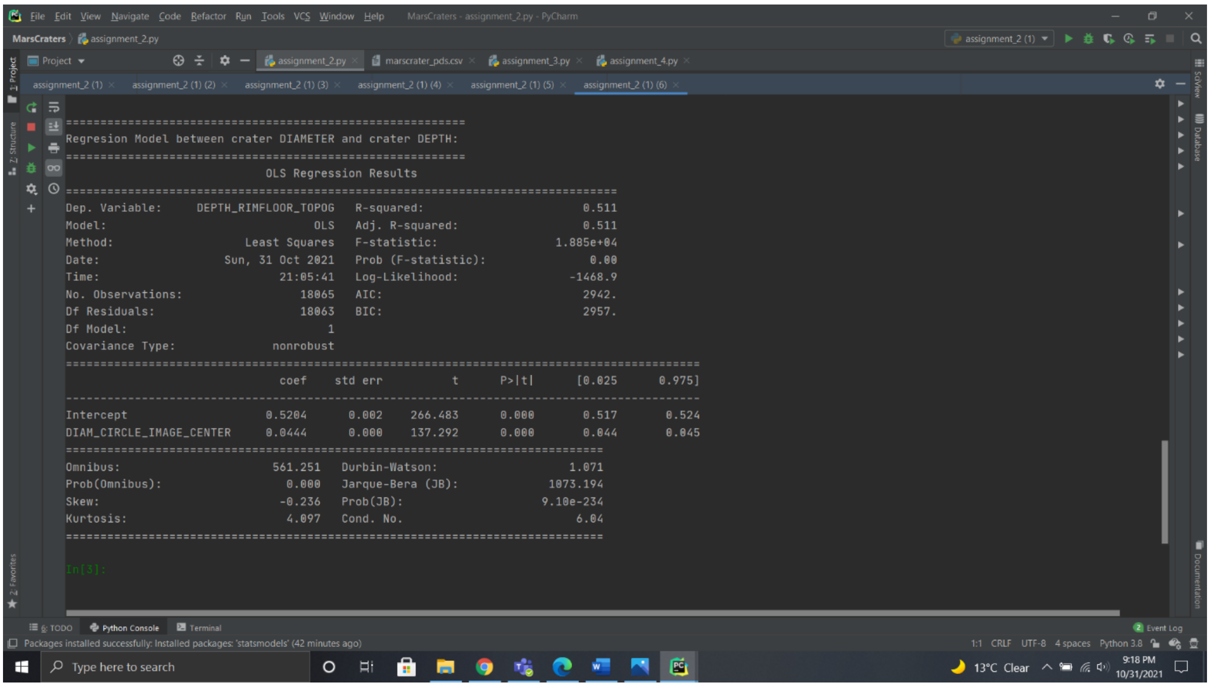
Interpretation of results:



```
=====================================
Statistics for Explanatory Variable:
=====================================

       DIAM_CIRCLE_IMAGE_CENTER

count              18065.000000
mean                   0.000000
std                    6.035266
min                   -7.175785
25%                   -3.705785
50%                   -1.835785
75%                    1.424215
max                   73.744215
```

Figure 2: Explanatory variable statistics

The table above shows the descriptive statistics of the centralized explanatory variable (DIAM_CIRCLE_IMAGE_CENTER). There is a zero mean indicating that the variable was properly concentrated.



```
==========================================================
Regresion Model between crater DIAMETER and crater DEPTH:
==========================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:     DEPTH_RIMFLOOR_TOPOG   R-squared:                       0.511
Model:                              OLS   Adj. R-squared:                  0.511
Method:                   Least Squares   F-statistic:                 1.885e+04
Date:                  Sun, 31 Oct 2021   Prob (F-statistic):               0.00
Time:                          21:05:41   Log-Likelihood:                -1468.9
No. Observations:                 18065   AIC:                             2942.
Df Residuals:                     18063   BIC:                             2957.
Df Model:                             1
Covariance Type:              nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                    0.5204      0.002    266.483      0.000       0.517       0.524
DIAM_CIRCLE_IMAGE_CENTER     0.0444      0.000    137.292      0.000       0.044       0.045
==============================================================================
Omnibus:                      561.251   Durbin-Watson:                   1.071
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1073.194
Skew:                          -0.236   Prob(JB):                    9.10e-234
Kurtosis:                       4.097   Cond. No.                         6.04
==============================================================================

In[3]:
```

Packages installed successfully: Installed packages: 'statsmodels' (42 minutes ago)

Figure 3: Explanatory variable results

The results of the linear regression model show that the diameter (Beta = 0.0444, p <0.0001) is significantly related to the crater depth. The table above shows a high F statistic (1,885 e + 04). A very low value p <0.0001 indicates that the relationship of crater diameter to crater depth is statistically significant. The regression model has an interruption of 0.5204 and a slope (regression coefficient) of 0.0444. An R-square of 0.511 suggests that if we know the crater diameter, we can predict 51% of the crater depth variability, while 49% variability will not be calculated. This means that we can predict over half the variability.

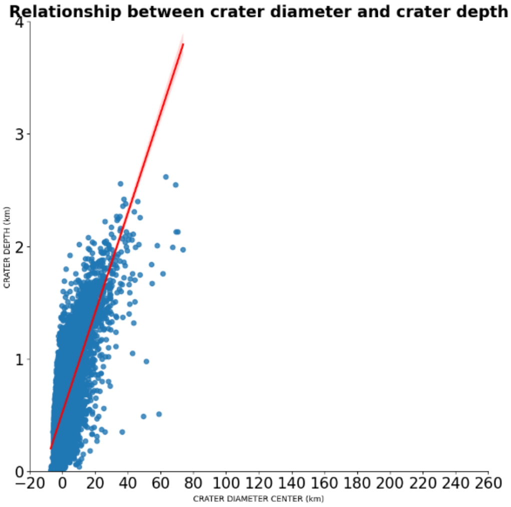


Figure 4: Relationship between crater diameter and depth.

The distribution graph above shows a positive linear relationship between crater diameter and crater depth; the greater crater depth is associated with higher crater diameters. The line of best fit is represented by the equation below;

$$crater\_depth \ = \ 0{,}044 \times crater\_diameter \ + \ 0{,}520$$

```
# ASSIGNMENT 2 for Course

data ['NUMBER_LAYERS'] = pd.to_numeric (data ['NUMBER_LAYERS'])
data ['DIAM_CIRCLE_IMAGE'] = pd.to_numeric (data
['DIAM_CIRCLE_IMAGE'])

data ['DEPTH_RIMFLOOR_TOPOG'] = pd.to_numeric (data
['DEPTH_RIMFLOOR_TOPOG'], errors = 'coerce')

# Frequency Distribution for NUMBER_LAWYERS Variable
print ()
print ('=========================================')
print ('Value counts for variable - NUMBER_LAYERS:')
print ('=========================================')

count_number_lawyers = data ['NUMBER_LAYERS']. value_counts (sort =
False)
print (count_number_lawyers)
print ()

# Percentages for NUMBER_LAWYERS Variable
print ()
print ('========================================')
print ('Percentages for variable - NUMBER_LAYERS:')
print ('========================================')

percentages_number_lawyers = data ['NUMBER_LAYERS']. value_counts
(sort = False, normalize = True)
print (count_number_lawyers)
print ()



# Frequency results
print ()
print ('=========')
print ('Results:')
print ('=========')

data_frame_1 = pd.DataFrame (np.array (count_number_lawyers), index
= count_number_lawyers.index, columns = ['Frequency'])

data_frame_2 = pd.DataFrame (np.array (percentages_number_lawyers *
100), index = count_number_lawyers.index, columns = ['Percentage'])
```

```
output_number_lawyers = pd.concat ([data_frame_1, data_frame_2],
axis = 1)

output_number_lawyers ['Frequency'] =
output_number_lawyers.Frequency.cumsum ()

output_number_lawyers ['Percentage'] =
output_number_lawyers.Percentage.cumsum ()

output_number_lawyers.index.name = 'NUMBER_LAWYERS'
print (output_number_lawyers)
print ()
```

The same was done for 2 other variables, respectively MORPHOLOGY_EJECTA_3 and MORPHOLOGY_EJECTA_2. The results are presented in a workbook created as an excel file.

7. Task 2: Making decisions about managed data

This part of the task involves making decisions about the data managed for the selected variables to answer the research questions. These data management decisions include removing invalid data, selecting the representative data set, creating secondary variables, and merging or grouping continuous variables.

From the data results we have 384,343 craters in the Mars database. From some data we have that 10 craters have negative depths, while 307,529 craters have depths equal to zero. This vague data of crater depth edge effects. Therefore, all craters (307,539) with negative depths and equal depths can be considered invalid. Of the total of 384,343 craters studied, only 76,512 craters were selected as valid data in the new data set we will be working with. Comparing the histogram distribution of the original data set of the dataset and the data set considered and selected as valid, they look very similar confirming that the selected data set is a representative "sample" of the entire dataset. . Therefore, any conclusions drawn about the data set we will be working with will be valid for the entire dataset obtained at the beginning of this task.

Using the new data set we will be working with, we created 3 new variables named respectively (DEPTH_GROUP, DIAMETER_GROUP and LATITUDE_GROUP) by merging the initial variables. Frequency distribution tables are constructed for each of the variables mentioned above.

```
# frequency distribution for variable LATITUDE_GROUP
latitude = [i for i in range (-90, 105, 15)]
latitude_labels = ['{0} - {1}'. format (i, i + 15) for i in range (-
90, 90, 15)]

working_subset ['LATITUDE_GROUP'] = pandas.cut
(working_subset.LATITUDE_CIRCLE_IMAGE, bins = latitude, labels =
latitude_labels, right = True)
```

```
print ()
print ('==========================================')
print ('Value counts for varaible - LATITUDE_GROUP:')
print ('==========================================')

count_latitude_group = working_subset ['LATITUDE_GROUP'].
value_counts (sort = False)

print (count_latitude_group)
print ()

# Percentages for LATITUDE_GROUP Variable
print ()
print ('=========================================')
print ('Percentages for variable - LATITUDE_GROUP:')
print ('=========================================')

percentages_latitude_group = working_subset ['LATITUDE_GROUP'].
value_counts (sort = False, normalize = True)

print (percentages_latitude_group)
print ()

print ()
print ('=========')
print ('Results:')
print ('=========')

data_frame_5 = pandas.DataFrame (numpy.array (count_latitude_group),
index = count_latitude_group.index.values, columns = ['Frequency'])

data_frame_6 = pandas.DataFrame (numpy.array
(percentages_latitude_group * 100), index =
percentages_latitude_group.index.values, columns = ['Percentage'])

output_latitude_group = pandas.concat ([data_frame_5, data_frame_6],
axis = 1)

output_latitude_group ['Frequency'] =
output_latitude_group.Frequency.cumsum ()

output_latitude_group ['Percentage'] =
output_latitude_group.Percentage.cumsum ()

output_latitude_group.index.name = 'LATITUDE_GROUP'

print (output_latitude_group)
print ()

# creating an workbook
excel_writer = pandas.ExcelWriter
('Mars_Craters_WorkBook_Assignment_3.xlsx')

output_depth_group.to_excel (excel_writer, sheet_name =
'DEPTH_GROUP', float_format = '% 0.4f', startrow = 1, startcol = 0)
```

```
output_diameter_group.to_excel (excel_writer, sheet_name =
'DIAMETER_GROUP', float_format = '% 0.4f', startrow = 1, startcol =
0)
```

```
output_latitude_group.to_excel (excel_writer, sheet_name =
'LATITUDE_GROUP', float_format = '% 0.4f', startrow = 1, startcol =
0)
```

```
excel_writer.save ()
```

```
wb = openpyxl.load_workbook
('Mars_Craters_WorkBook_Assignment_3.xlsx')
```

```
sheets = wb.sheetnames
```

```
for i in range (len (sheets)):
 ws = wb.worksheets [i]
 variable_name = ws.cell (row = 2, column = 1) .value
 title = 'Mars Crater Dataset Frequency Distribution for variable:'
+ variable_name

 ws.cell (row = 1, column = 1) .value = title.upper ()
 ws.merge_cells (start_row = 1, start_column = 1, end_row = 1,
end_column = 5)
 wb.save ('Mars_Craters_WorkBook_Assignment_3.xlsx')
```

The same was done for 2 other variables, respectively DEPTH_GROUP and DIAMETER_GROUP. The results are presented in a workbook created as an excel file.

NUMBER_LAWYERS	Frequency	Percentage
0	364612	94.8663
1	15467	4.0243
2	3435	0.8937
3	739	0.1923
4	85	0.0221
5	5	0.0013

Table 2: Statistics table for NUMBER_LAWYERS

Interpretation:

The Mars Crater Database contains a record of the number of layers each crater has. This record is represented by the variable NUMBER_LAYERS. Of the total 384,343 craters studied, about 94.87% have no ejection layer, 15,467 craters have 1 ejection layer, comprising 4.02% of the data, 0.89% and 0.19% of the craters have respectively 2 to 3 number of ejection layers. The largest number of extraction layer is 5, where only 5 Martian craters (0.001%) are included in this category.

The craters were divided into 3 morphological groups based on the shape and composition of the crater, named respectively with the variables: MORPHOLOGY_EJECTA_1, MORPHOLOGY_EJECTA_2 and MORPHOLOGY_EJECTA_3. Each group was further divided into different morphological classes. The second variable to be studied is the group

MORPHOLOGY_EJECTA_3, which contains 28 individual morphological classifications and the third variable group MORPHOLOGY_EJECTA_2, which contains 103 individual morphological classifications. For these we have created separate workbooks to display the results obtained.

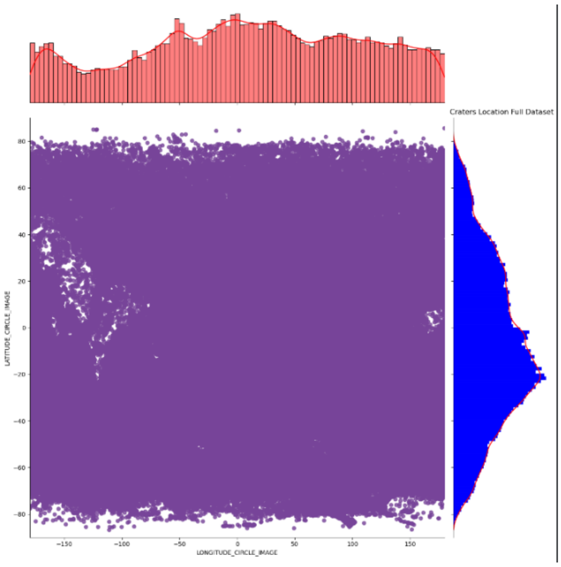


Figure 5: Distribution of location points of the subset.

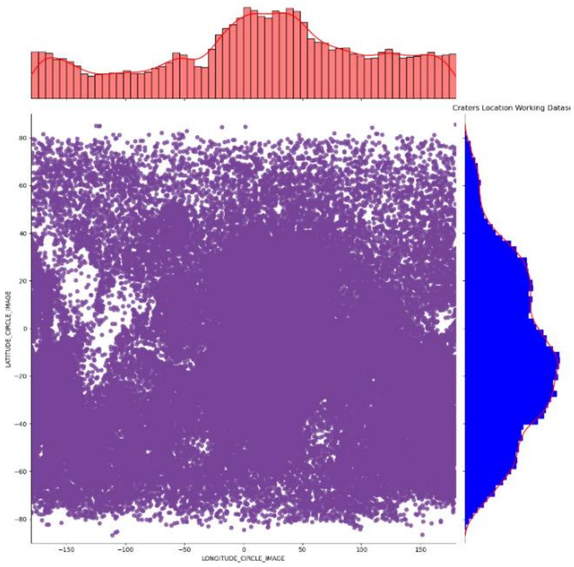


Figure 6: Distribution of dataset location points.

The graph of the above points (Figure 5) shows a 2D map of the entire original crater database. All 384,343 craters were plotted with latitude on the Y axis and longitude on the X axis. The red histogram at the top of the graph shows the crater length distribution while the blue histogram to the right of the graph shows the crater width distribution. We see that the number of craters is too large for any visualization of the points of their location. Defining a smaller data set is necessary for an optimal data analysis.

The second graph (Figure 6) is a distribution graph of the diameter (sizes) of the crater versus the depth of the crater. This schema was used to define a new data set with all invalid points removed. The new data set includes only craters with a diameter greater than 0 and less than or equal to 100 km; and depth greater than 0 and less than or equal to 3 km.

We create a new variable called DEPTH_GROUP by grouping the crater depth variable (DEPTH_RIMFLOOR_TOPOG) into 10 different groups. We also create a second new variable named DIAMETER_GROUP to group the crater diameter variable (DIAM_CIRCLE_IMAGE) into 20 different groups. A third and final new variable, LATITUDE_GROUP, is created by grouping the latitude coordinate variable (LATITUDE_CIRCLE_IMAGE) into 12 different groups or regions.

The results are presented in a separate workbook created as an excel file.

8. Task 3: Create graphs to visualize data

This part of the task is about providing a visual representation of the data using graphs. Using different types of data visualization, we will answer some of the research questions listed below:

- Is the diameter of the crater related to the depth of the crater?
- Do craters occur most often near the equator, north or south?
- Does the depth of the crater depend on the location of the crater?

Recalling the graph of task 2 (Figure 6) is a 2D map of the crater location showing all the craters (76,512) in the data set we got to study. This data set includes only craters with a diameter greater than 0 and less than or equal to 100 km; and depth greater than 0 and less than or equal to 3 km.

The crater location graph reveals a general trend in crater appearance with some well-defined crater clusters. In general, there are more craters in the southern hemisphere (0 to -90 degrees latitude) than in the northern hemisphere (0 to 90 degrees latitude). Furthermore, the density of craters increases in the equatorial zone of Mars (-45 to 45 degrees latitude), but decreases towards the edges of both the north and south poles. The red histogram at the beginning of the graph shows the crater length distribution while the blue histogram to the right of the graph shows the crater width distribution.

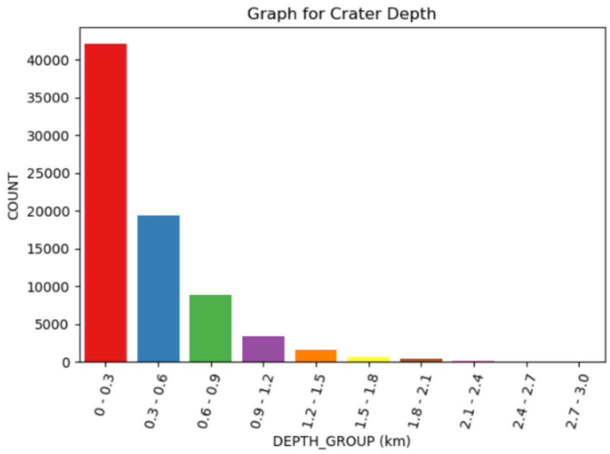


Figure 7: Graph for the DEPTH_GROUP variable

This graph is a 1-modal graph (1 maximum) with its highest peak in category 0 to 0.3 km depth. The graph is skewed to the right as it has higher frequencies in the lower crater depth groups. Shallow craters are more numerous than larger depth craters on Mars.

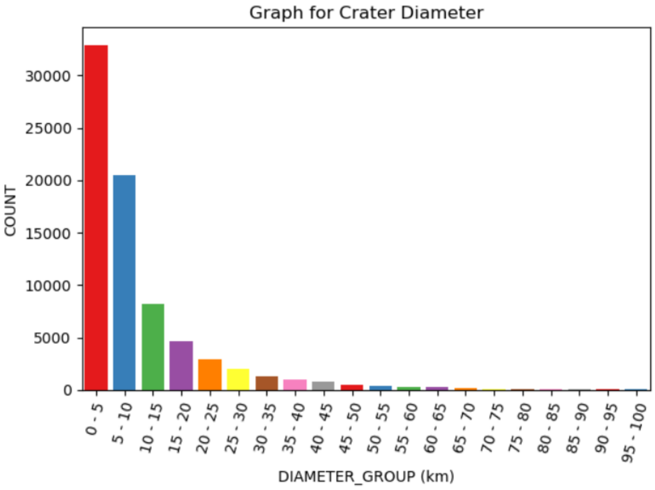


Figure 8: Graph for the variable DIAMETER_GROUP

This graph is also 1-modal with its highest peak in the 0 to 5 km diameter category. It is sloping to the right as it has higher frequencies in the groups with the lowest crater diameter. Smaller (narrower) craters are more numerous than large (wider) craters.

The following graph is 1-modal with its highest peak in the category -30 to -15 degrees latitude. The graph is skewed to the right, but is not highlighted. It also appears to have a semi-symmetrical shape. This suggests that there are more craters near the equatorial region of Mars (-45 to 45 degrees latitude) than at the North Pole (45 to 90 degrees latitude) and the South Pole (-45 to -90 degrees latitude).

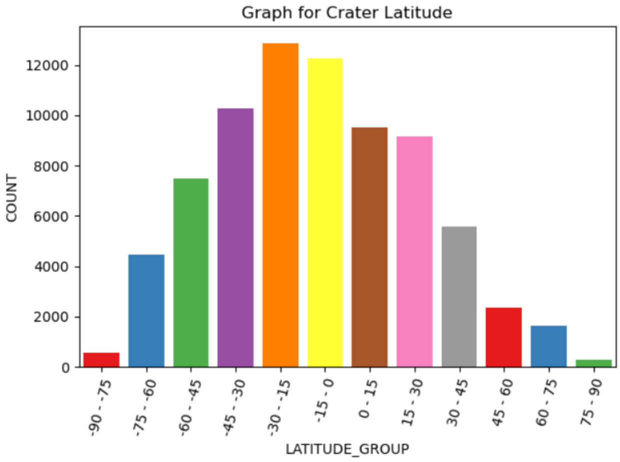


Figure 9: Graph for the variable LATITUDE_GROUP

To answer our questions, which we posed above, we also did a visualization of the data to get an answer as accurate as possible.

1. Is there a correlation between the diameter and depth of the crater?

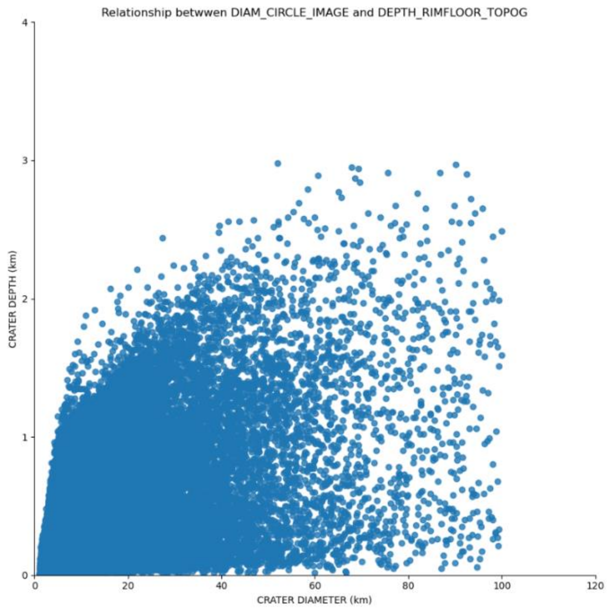


Figure 10: Relationship between crater diameter and depth

The graph above shows the depth of craters in the diameter of craters. We can see that the distribution graph appears centered and shows a relationship or trend between the two variables. We can not say for sure, but we can see a strong linear relationship between the average depth of the crater and the average diameter of the crater, given that the concentration of points is in a linear area of the graph.

```
working_subset ['LATITUDE_GROUP'] = working_subset
['LATITUDE_GROUP']. astype ('category')

graph_5 = seaborn.countplot (
 x = 'LATITUDE_GROUP', data = working_subset, saturation = 1,
palette = 'Set1'
)

plt.xlabel ('LATITUDE_GROUP')
plt.ylabel ('COUNT')
plt.title ('Graph for Crater Latitude')
plt.tick_params (axis = 'both')
plt.xticks (rotation = 75)
plt.show ()

working_subset ['DIAMETER_GROUP'] = working_subset
['DIAMETER_GROUP']. astype ('category')

graph_4 = seaborn.countplot (
 x = 'DIAMETER_GROUP', data = working_subset, saturation = 1,
palette = 'Set1'
)
```

```
plt.xlabel ('DIAMETER_GROUP (km)')
plt.ylabel ('COUNT')
plt.title ('Graph for Crater Diameter')
plt.tick_params (axis = 'both')
plt.xticks (rotation = 75)
plt.show ()

graph_3 = seaborn.countplot (
 x = 'DEPTH_GROUP', data = working_subset, saturation = 1, palette =
'Set1'
)

plt.xlabel ('nazi5tamiNic1 (km)')
plt.ylabel ('COUNT')
plt.title ('Graph for Crater Depth')
plt.tick_params (axis = 'both')
plt.xticks (rotation = 75)
plt.show ()
```

References:

[1] Barlow, Nadine. "Impact craters in the northern hemisphere of Mars: Layer ejecta and central pit characteristics". Meteoritics & Planetary Science 41, No. 10, (2006): 1425–1436.

[2] Barlow NG and Perez CB 2003. Martian impact crater ejecta morphologies as indicators of the distribution of volatile subsurface. Journal of Geophysical Research, doi: 10.1029 / 2002JE002036.

[3] Robbins, Stuart. "Planetary Surface Properties, Cratering Physics, and the Volcanic History of Mars from a New Global Martian Crater Database." PhD Thesis, University of Colorado (2011): 251 pages.

[4] http://about.sjrdesign.net/files/thesis/RobbinsThesis_LargeMB.pdf;

[5] https://www.coursera.org/learn/data-visualization/home/welcome

Mexhit Kurti
Faculty of Natural Sciences
Scientific Master in Informatics
Data Mining
2021