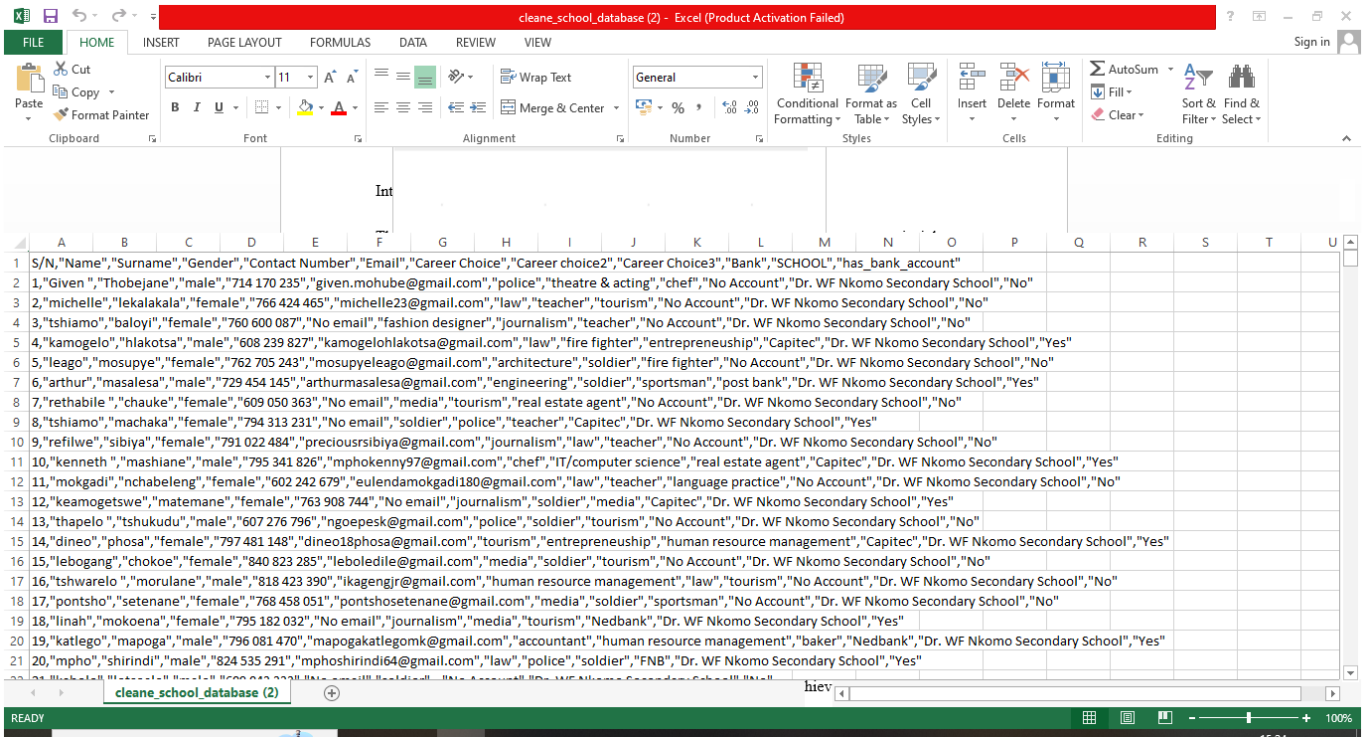
Report: Data Cleaning and Exploration of Student Banking Dataset

Introduction

This project involved cleaning and analyzing a student banking dataset to uncover insights about student banking behavior and prepare the data for further analysis or modeling. The dataset included various fields such as student personal details, career interests, and bank information. The main objective was to ensure data quality and consistency—especially around bank usage—and to generate meaningful insights that can support financial inclusion strategies, partnerships with banks, and career-based financial profiling.

RAW DATA



```
S/N,"Name","Surname","Gender","Contact Number","Email","Career Choice","Career choice2","Career Choice3","Bank","SCHOOL","has_bank_account"
1,"Given ","Thobejane","male","714 170 235","given.mohube@gmail.com","police","theatre & acting","chef","No Account","Dr. WF Nkomo Secondary School","No"
2,"michelle","lekalakala","female","766 424 465","michelle23@gmail.com","law","teacher","tourism","No Account","Dr. WF Nkomo Secondary School","No"
3,"tshiamo","baloyi","female","760 600 087","No email","fashion designer","journalism","teacher","No Account","Dr. WF Nkomo Secondary School","No"
4,"kamogelo","hlakotsa","male","608 239 827","kamogelohlakotsa@gmail.com","law","fire fighter","entrepreneuship","Capitec","Dr. WF Nkomo Secondary School","Yes"
5,"leago","mosupye","female","762 705 243","mosupyeleago@gmail.com","architecture","soldier","fire fighter","No Account","Dr. WF Nkomo Secondary School","No"
6,"arthur","masalesa","male","729 454 145","arthurmasalesa@gmail.com","engineering","soldier","sportsman","post bank","Dr. WF Nkomo Secondary School","Yes"
7,"rethabile ","chauke","female","609 050 363","No email","media","tourism","real estate agent","No Account","Dr. WF Nkomo Secondary School","No"
8,"tshiamo","machaka","female","794 313 231","No email","soldier","police","teacher","Capitec","Dr. WF Nkomo Secondary School","Yes"
9,"refilwe","sibiya","female","791 022 484","preciousrsibiya@gmail.com","journalism","law","teacher","No Account","Dr. WF Nkomo Secondary School","No"
10,"kenneth ","mashiane","male","795 341 826","mphokenny97@gmail.com","chef","IT/computer science","real estate agent","Capitec","Dr. WF Nkomo Secondary School","Yes"
11,"mokgadi","nchabeleng","female","602 242 679","eulendamokgadi180@gmail.com","law","teacher","language practice","No Account","Dr. WF Nkomo Secondary School","No"
12,"keamogetswe","matemane","female","763 908 744","No email","journalism","soldier","media","Capitec","Dr. WF Nkomo Secondary School","Yes"
13,"thapelo ","tshukudu","male","607 276 796","ngoepesk@gmail.com","police","soldier","tourism","No Account","Dr. WF Nkomo Secondary School","No"
14,"dineo","phosa","female","797 481 148","dineo18phosa@gmail.com","tourism","entrepreneuship","human resource management","Capitec","Dr. WF Nkomo Secondary School","Yes"
15,"lebogang","chokoe","female","840 823 285","leboledile@gmail.com","media","soldier","tourism","No Account","Dr. WF Nkomo Secondary School","No"
16,"tshwarelo ","morulane","male","818 423 390","ikagengjr@gmail.com","human resource management","law","tourism","No Account","Dr. WF Nkomo Secondary School","No"
17,"pontsho","setenane","female","768 458 051","pontshosetenane@gmail.com","media","soldier","sportsman","No Account","Dr. WF Nkomo Secondary School","No"
18,"linah","mokoena","female","795 182 032","No email","journalism","media","tourism","Nedbank","Dr. WF Nkomo Secondary School","Yes"
19,"katlego","mapoga","male","796 081 470","mapogakatlegomk@gmail.com","accountant","human resource management","baker","Nedbank","Dr. WF Nkomo Secondary School","Yes"
20,"mpho","shirindi","male","824 535 291","mphoshirindi64@gmail.com","law","police","soldier","FNB","Dr. WF Nkomo Secondary School","Yes"
```

Data Cleaning Process Using SQL

To begin, a comprehensive SQL cleaning script was developed to refine the students table, focusing particularly on the bank and has_bank_account columns. The first step involved **removing duplicate rows** to prevent skewed statistics. This was achieved by grouping similar records and retaining only the first occurrence based on multiple identifying fields such as name, surname, school, and bank.

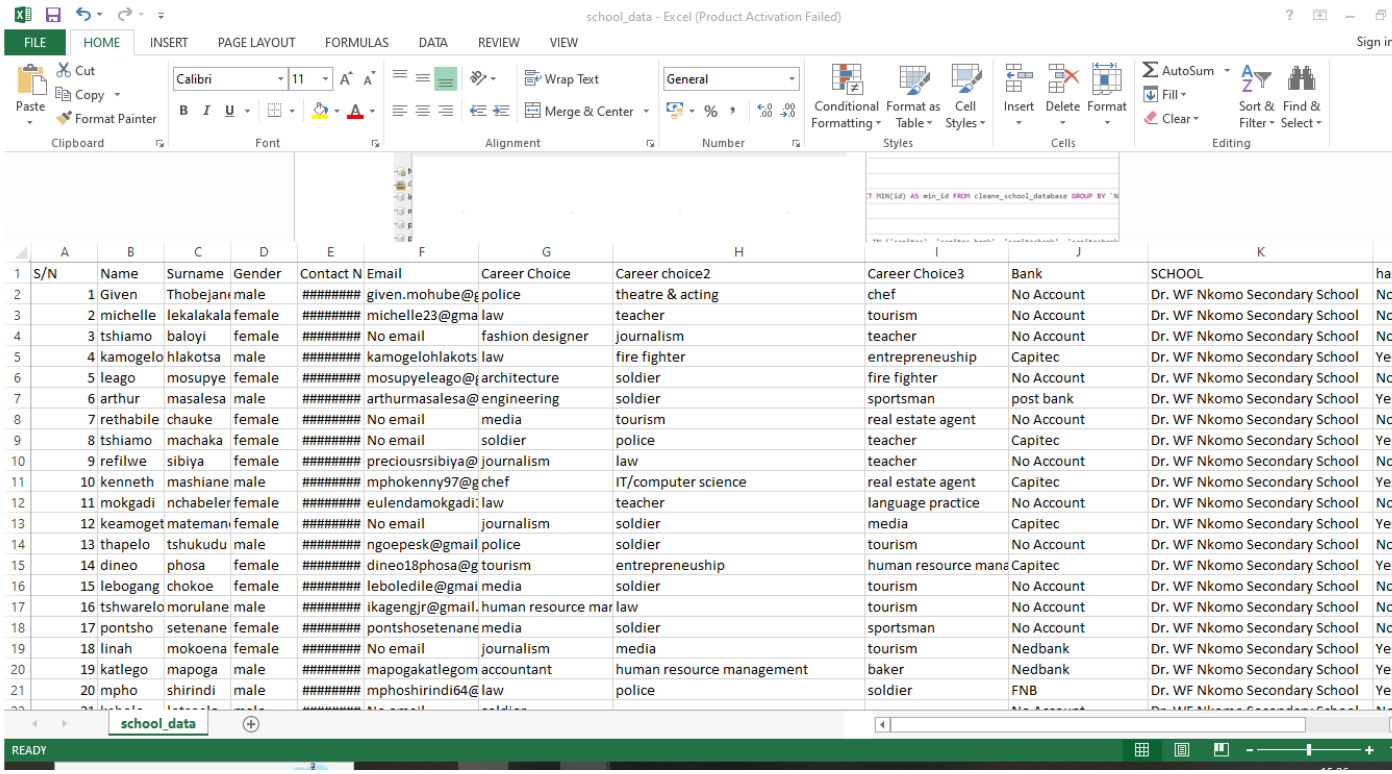
The second step addressed **inconsistent formatting**, particularly trimming any leading or trailing spaces around bank names using the TRIM() function. Following that, **bank name standardization** was conducted to consolidate variations of the same bank—for example, transforming entries like "capitec bank" and "capitecbank" into "Capitec", and similarly resolving entries for ABSA, FNB, Nedbank, and Standard Bank. These updates helped reduce redundancy and confusion in the dataset.

After standardizing the known banks, the next cleaning phase involved identifying and labeling **missing or empty bank names**. These were replaced with a uniform label "No Account", representing students who do not currently hold a bank account. This helped isolate financially excluded students and set the foundation for further analysis or modeling.

Optionally, special characters within bank names were considered for removal using REGEXP_REPLACE, although this step depends on database support for regular expressions. Additionally, the bank names were normalized into **lowercase** for consistency across queries and analytics. Some systems could also support conversion into proper case (INITCAP()), if a more human-readable format is desired.

As a final cleanup step, a **summary view** was created to report on the total number of students per bank. This summary can be used for dashboards, quick analysis, or integration into BI tools. An additional column called has_bank_flag was also introduced to clearly distinguish students with and without accounts using a Boolean value (TRUE for account holders, FALSE for non-holders).

CLEAN DATA



	A	B	C	D	E	F	G	H	I	J	K
1	S/N	Name	Surname	Gender	Contact N	Email	Career Choice	Career choice2	Career Choice3	Bank	SCHOOL
2	1	Given	Thobejan	male	########	given.mohube@g	police	theatre & acting	chef	No Account	Dr. WF Nkomo Secondary School
3	2	michelle	lekalakala	female	########	michelle23@gma	law	teacher	tourism	No Account	Dr. WF Nkomo Secondary School
4	3	tshiamo	baloyi	female	########	No email	fashion designer	journalism	teacher	No Account	Dr. WF Nkomo Secondary School
5	4	kamogelo	hlakotsa	male	########	kamogelohlakots	law	fire fighter	entrepreneuship	Capitec	Dr. WF Nkomo Secondary School
6	5	leago	mosupye	female	########	mosupyeleago@g	architecture	soldier	fire fighter	No Account	Dr. WF Nkomo Secondary School
7	6	arthur	masalesa	male	########	arthurmasalesa@	engineering	soldier	sportsman	post bank	Dr. WF Nkomo Secondary School
8	7	rethabile	chauke	female	########	No email	media	tourism	real estate agent	No Account	Dr. WF Nkomo Secondary School
9	8	tshiamo	machaka	female	########	No email	soldier	police	teacher	Capitec	Dr. WF Nkomo Secondary School
10	9	refilwe	sibiya	female	########	preciousrsibiya@	journalism	law	teacher	No Account	Dr. WF Nkomo Secondary School
11	10	kenneth	mashiane	male	########	mphokenny97@g	chef	IT/computer science	real estate agent	Capitec	Dr. WF Nkomo Secondary School
12	11	mokgadi	nchabeler	female	########	eulendamokgadi	law	teacher	language practice	No Account	Dr. WF Nkomo Secondary School
13	12	keamoget	mateman	female	########	No email	journalism	soldier	media	Capitec	Dr. WF Nkomo Secondary School
14	13	thapelo	tshukudu	male	########	ngoepesk@gmail	police	soldier	tourism	No Account	Dr. WF Nkomo Secondary School
15	14	dineo	phosa	female	########	dineo18phosa@g	tourism	entrepreneuship	human resource mana	Capitec	Dr. WF Nkomo Secondary School
16	15	lebogang	chokoe	female	########	leboledile@gmai	media	soldier	tourism	No Account	Dr. WF Nkomo Secondary School
17	16	tshwarelo	morulane	male	########	ikagengjr@gmail.	human resource mar	law	tourism	No Account	Dr. WF Nkomo Secondary School
18	17	pontsho	setenane	female	########	pontshosetenane	media	soldier	sportsman	No Account	Dr. WF Nkomo Secondary School
19	18	linah	mokoena	female	########	No email	journalism	media	tourism	Nedbank	Dr. WF Nkomo Secondary School
20	19	katlego	mapoga	male	########	mapogakatlegom	accountant	human resource management	baker	Nedbank	Dr. WF Nkomo Secondary School
21	20	mpho	shirindi	male	########	mphoshirindi64@	law	police	soldier	FNB	Dr. WF Nkomo Secondary School

Analysis Insights (from Python & SQL Integration)

Using Python (with Pandas and Matplotlib), the cleaned data was further analyzed. The top 5 most commonly used banks among students were identified and visualized. Capitec, ABSA, and FNB emerged as the most popular institutions, indicating strong brand presence and adoption among learners.

Students without bank accounts were filtered and analyzed based on their **career interests** and **schools**. This revealed that certain career groups—such as those interested in the arts or education—had a higher proportion of unbanked individuals. Similarly, a few schools stood out for having significantly more students without accounts. These findings can help banks and school administrators target their outreach programs effectively.

lit View Run Kernel Settings Help Trusted

Code JupyterLab Python (Pyodide)

```
import pandas as pd
import numpy as np


from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score
import matplotlib.pyplot as plt


%pip install seaborn
import seaborn as sns
df = pd.read_csv("school_data.csv", delimiter=';')
df = df.drop(['Contact Number', 'Email','Surname'], axis=1)
df['Bank'] = df['Bank'].replace([None, 'none', 'None', '', 'nan', 'NaN'], 'No Account')

df.head(5)
```

	S/N	Name	Gender	Career Choice	Career choice2	Career Choice3	Bank	SCHOOL	has_bank_account
0	1	Given	male	police	theatre & acting	chef	No Account	Dr. WF Nkomo Secondary School	No
1	2	michelle	female	law	teacher	tourism	No Account	Dr. WF Nkomo Secondary School	No
2	3	tshiamo	female	fashion designer	journalism	teacher	No Account	Dr. WF Nkomo Secondary School	No
3	4	kamogelo	male	law	fire fighter	entrepreneuship	Capitec	Dr. WF Nkomo Secondary School	Yes
4	5	leago	female	architecture	soldier	fire fighter	No Account	Dr. WF Nkomo Secondary School	No

•[6]:

```
#how many people have bank account
bank_account_counts = df['has_bank_account'].value_counts()
print(bank_account_counts)
```

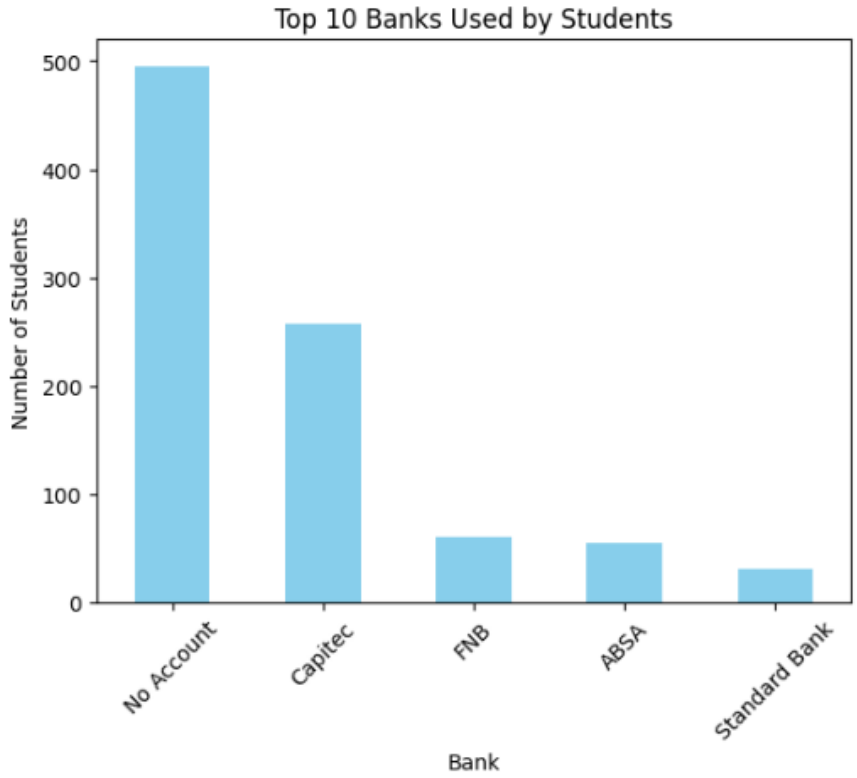
```
has_bank_account
Yes    600
No     333
Name: count, dtype: int64
```

•[11]:

```
#top bank used
top_banks = df['Bank'].value_counts().head(5)
print(top_banks)


top_banks.plot(kind='bar', color='skyblue')
plt.title("Top 10 Banks Used by Students")
plt.xlabel("Bank")
plt.ylabel("Number of Students")
plt.xticks(rotation=45)
plt.show()
```

```
Bank
No Account       495
Capitec          258
FNB               61
ABSA              55
Standard Bank     31
Name: count, dtype: int64
```



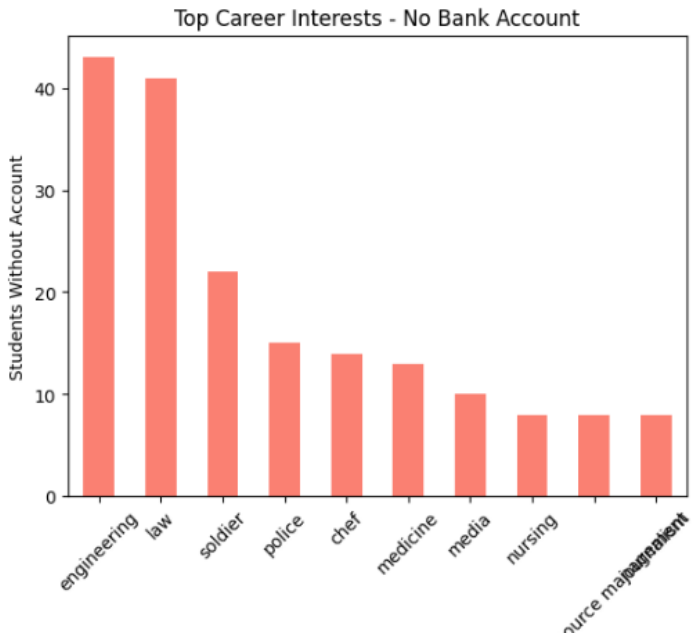
```python
no_account_df = df[df['has_bank_account'] == 'No']

# By career
print("Career Interests of Students WITHOUT Accounts:")
print(no_account_df['Career Choice'].value_counts().head(10))

# By school (optional)
print("\nSchools with Most Students Without Accounts:")
print(no_account_df['SCHOOL'].value_counts().head(10))

# Optional plot
no_account_df['Career Choice'].value_counts().head(10).plot(kind='bar', color='salmon')
plt.title("Top Career Interests - No Bank Account")
plt.xlabel("Career Choice")
plt.ylabel("Students Without Account")
plt.xticks(rotation=45)
plt.show()
```

```
Career Interests of Students WITHOUT Accounts:
Career Choice
engineering                 43
law                         41
soldier                     22
police                      15
chef                        14
medicine                    13
media                       10
nursing                      8
human resource management    8
journalism                   8
Name: count, dtype: int64
```



Uneven Access to Bank Accounts

Out of 933 students, 600 (64%) have bank accounts, while 333 (36%) do not. While the majority are banked, over one-third of students remain financially excluded. This is a significant number, considering the importance of banking in enabling students to receive bursaries, manage allowances, and build early financial responsibility. The gap suggests that **more targeted education or partnerships may be necessary** to improve access for all learners.

Bank Preferences: Capitec Dominates

Capitec emerged as the **most used bank** by students, with 258 users—more than four times the usage of FNB (61) and ABSA (55). This may reflect Capitec's accessible and affordable banking model, as well as its strong marketing presence in youth markets. The popularity of Capitec can guide banks looking to compete in the youth segment to offer **more simplified, mobile-friendly** account types.

Schools with Financial Exclusion

Some schools have disproportionately high numbers of unbanked students. Dr. WF Nkomo Secondary School, for instance, has 74 students without accounts—more than twice the next highest. This highlights the need for **school-level interventions**, such as:

- Financial literacy programs
- Bank account registration days
- Digital wallet alternatives

By identifying the specific schools most affected, banks or education departments can **deploy focused strategies** rather than broad, less efficient campaigns.

Career-Based Financial Gaps

Career interest also appears to influence banking access. Students interested in **engineering (43 unbanked)** and **law (41 unbanked)** lead the list of those without bank accounts. This suggests that even students pursuing professional and high-potential careers may face access issues. It may also reflect **socioeconomic factors**, where some career tracks are more common in under-resourced communities.

Banks can use this insight to design **career-aligned banking programs**—for example, offering early access to student loans or financial advice tailored to specific fields.

Conclusion

These findings can support initiatives that promote banking education, career-linked banking services, and targeted outreach programs in underserved schools. The clean and structured dataset can also serve as the foundation for machine learning, financial forecasting, and policy development focused on youth financial inclusion.