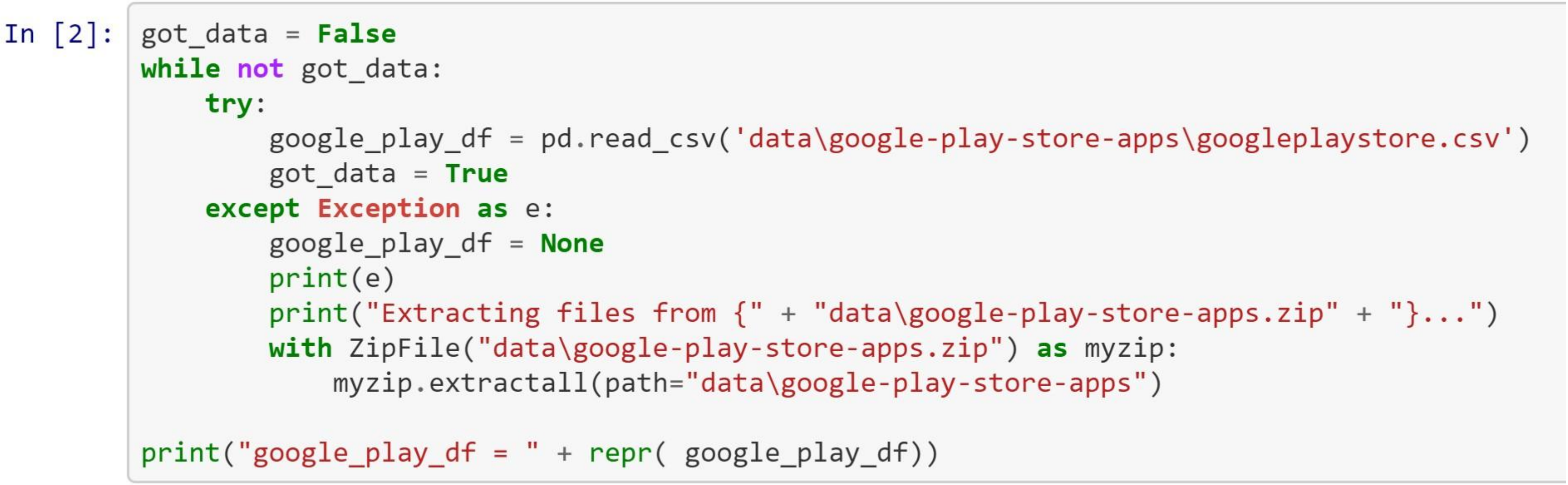
Case Study
August 28, 2020

Mike Collins
Andrew Perry

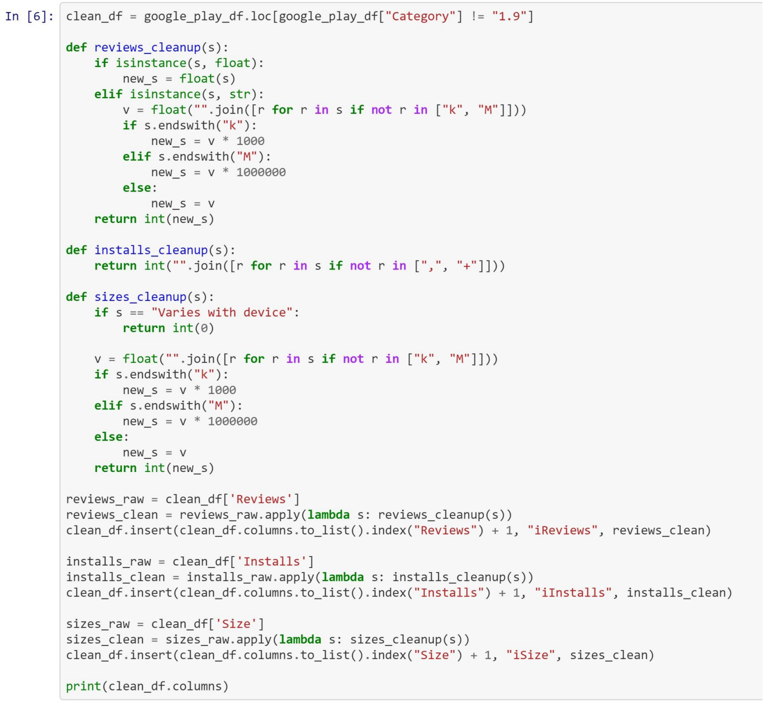
Importing the data from a zip file



```
In [2]: got_data = False
        while not got_data:
            try:
                google_play_df = pd.read_csv('data\google-play-store-apps\googleplaystore.csv')
                got_data = True
            except Exception as e:
                google_play_df = None
                print(e)
                print("Extracting files from {" + "data\google-play-store-apps.zip" + "}...")
                with ZipFile("data\google-play-store-apps.zip") as myzip:
                    myzip.extractall(path="data\google-play-store-apps")

        print("google_play_df = " + repr( google_play_df))
```

Cleaning up the data



In [6]:

```
clean_df = google_play_df.loc[google_play_df["Category"] != "1.9"]

def reviews_cleanup(s):
    if isinstance(s, float):
        new_s = float(s)
    elif isinstance(s, str):
        v = float("".join([r for r in s if not r in ["k", "M"]]))
        if s.endswith("k"):
            new_s = v * 1000
        elif s.endswith("M"):
            new_s = v * 1000000
        else:
            new_s = v
    return int(new_s)

def installs_cleanup(s):
    return int("".join([r for r in s if not r in [",", "+"]]))

def sizes_cleanup(s):
    if s == "Varies with device":
        return int(0)

    v = float("".join([r for r in s if not r in ["k", "M"]]))
    if s.endswith("k"):
        new_s = v * 1000
    elif s.endswith("M"):
        new_s = v * 1000000
    else:
        new_s = v
    return int(new_s)

reviews_raw = clean_df['Reviews']
reviews_clean = reviews_raw.apply(lambda s: reviews_cleanup(s))
clean_df.insert(clean_df.columns.to_list().index("Reviews") + 1, "iReviews", reviews_clean)

installs_raw = clean_df['Installs']
installs_clean = installs_raw.apply(lambda s: installs_cleanup(s))
clean_df.insert(clean_df.columns.to_list().index("Installs") + 1, "iInstalls", installs_clean)

sizes_raw = clean_df['Size']
sizes_clean = sizes_raw.apply(lambda s: sizes_cleanup(s))
clean_df.insert(clean_df.columns.to_list().index("Size") + 1, "iSize", sizes_clean)

print(clean_df.columns)
```

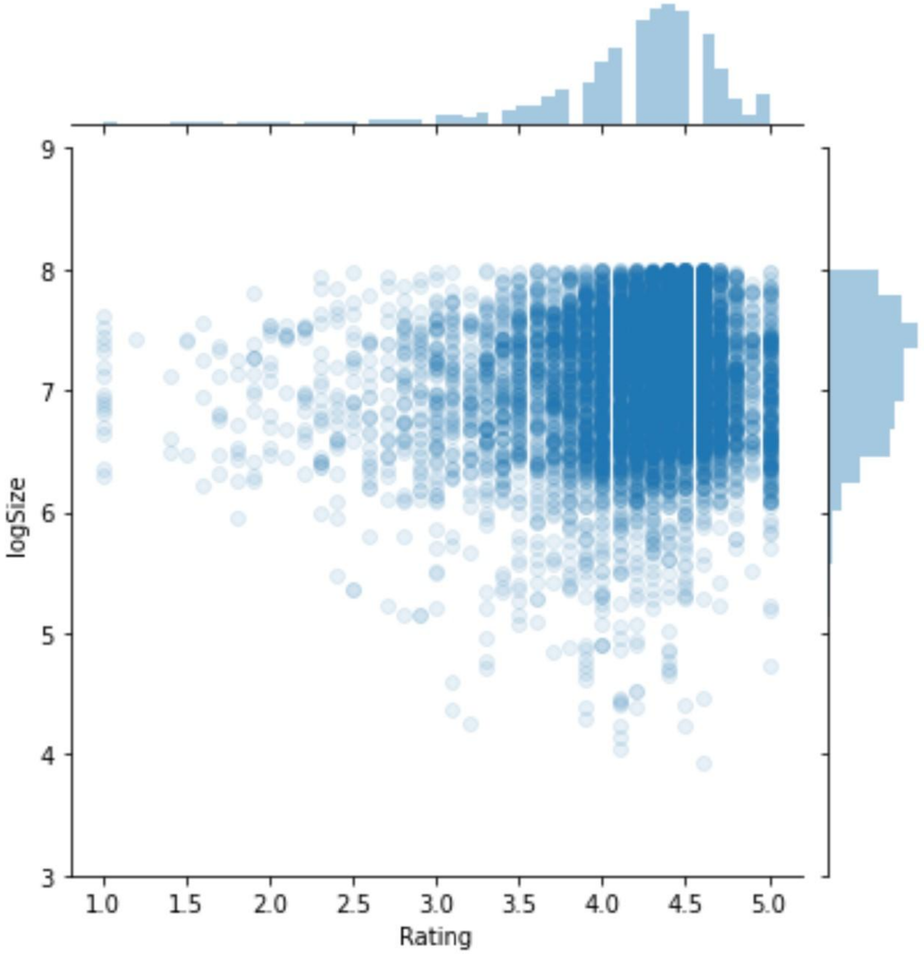
Graph for Question 2

Relationship Between

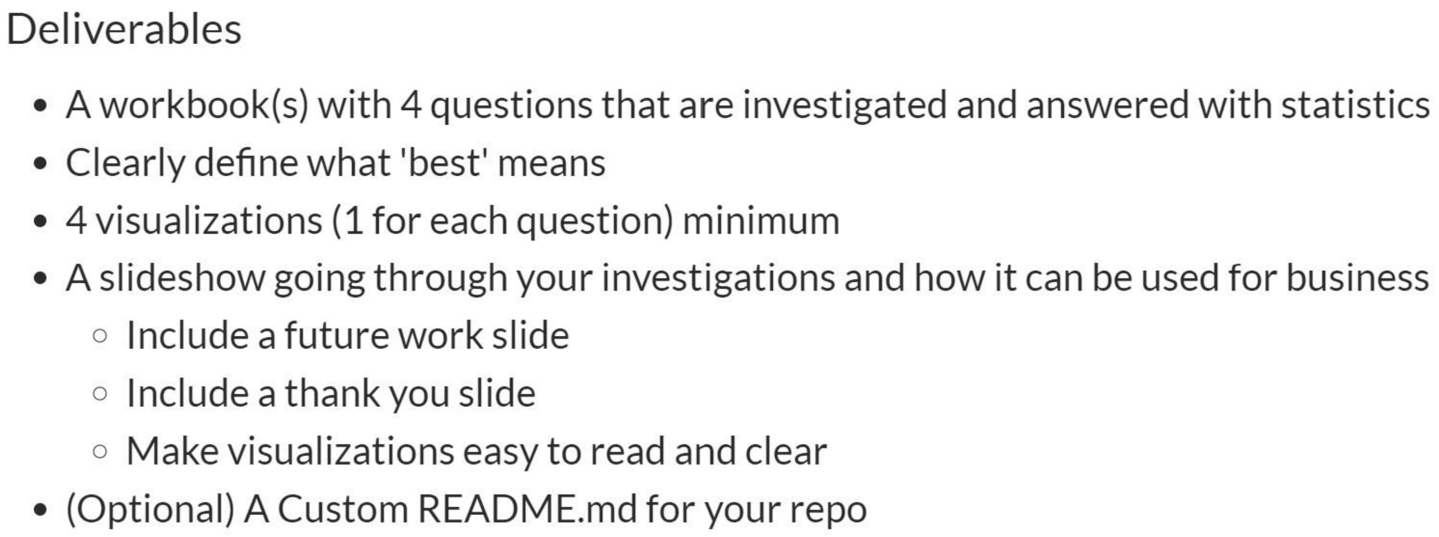
“Size of Download”

And

“Rating” of application



What we were supposed to do



Deliverables

- A workbook(s) with 4 questions that are investigated and answered with statistics
- Clearly define what 'best' means
- 4 visualizations (1 for each question) minimum
- A slideshow going through your investigations and how it can be used for business
    - Include a future work slide
    - Include a thank you slide
    - Make visualizations easy to read and clear
- (Optional) A Custom README.md for your repo

Future Work:

Do the actual work

Thank you: