Project 1 What is in a name

Austin Wilson

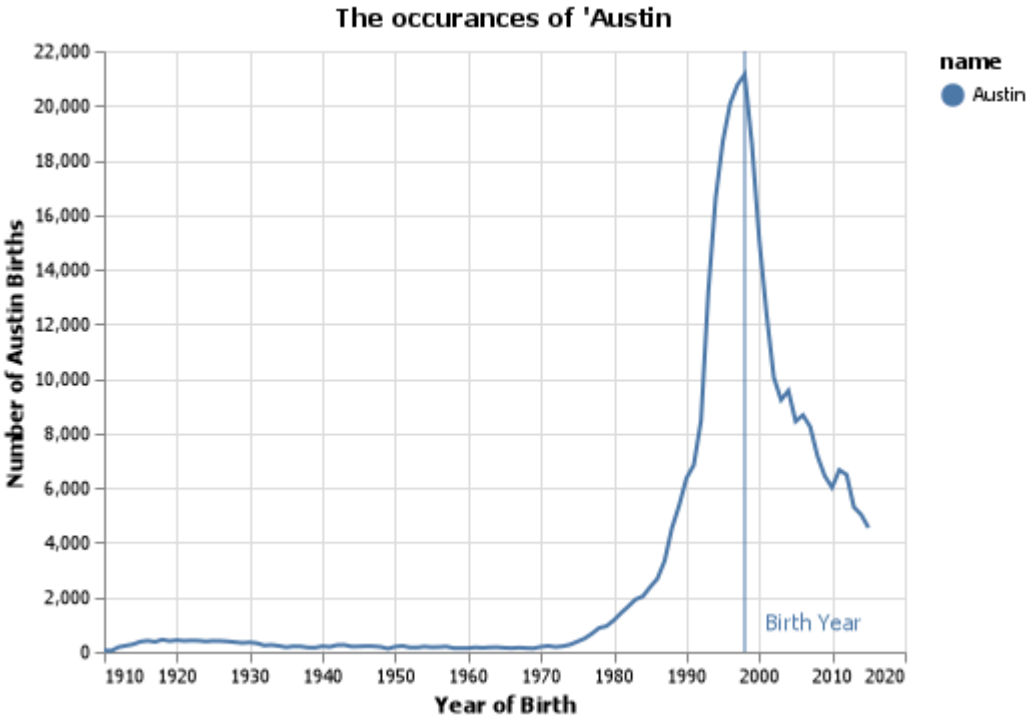
Elevator pitch

This project is focused around analyzing data about names and their occurance over the years. Using this data i was able to seperate and sort the data in different graphs to answer the question listed below. What i hope to show with this project is the ability to take large data sets and answer small and specific questions with data and with visuals.

TECHNICAL DETAILS

GRAND QUESTION 1

How does your name at your birth year compare to its use historically?

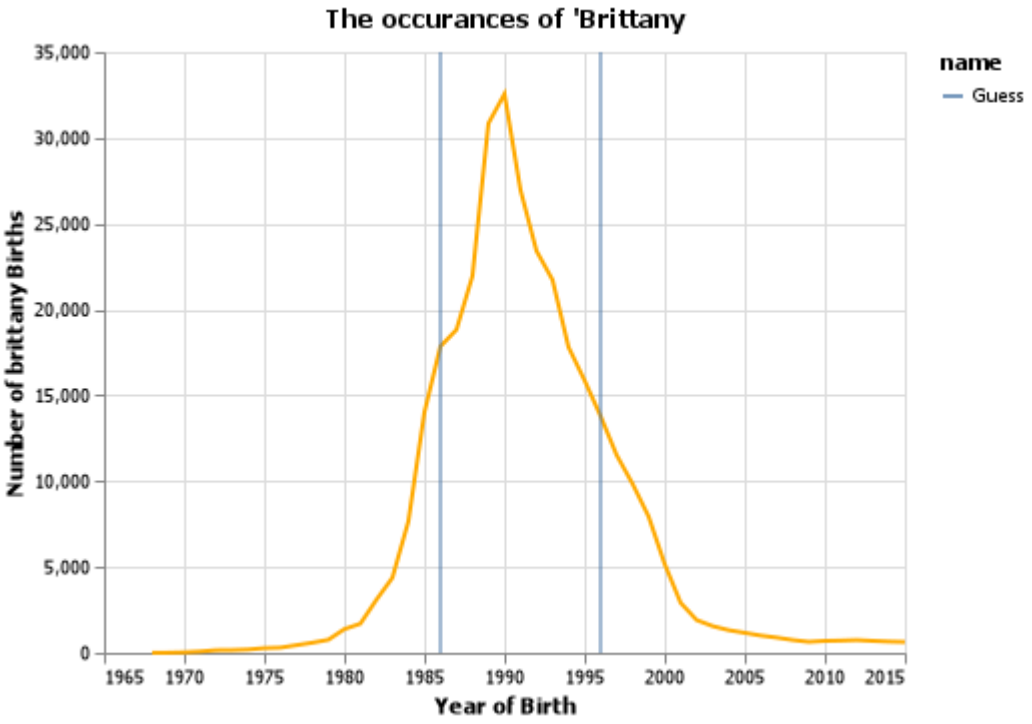


At my birth year of 1998 what we see from the data is a spike to record high numbers, followed by decreasing occurances over time. I think that what this shows us is an anomaly of increasing popularity from 1985 to about 1997. I suspect this spike is due to the movie Austin Powers, but i cannot prove it without more data.

GRAND QUESTION 2

If you talked to someone named Brittany on the phone, what is your guess of their age? What ages would you not guess?

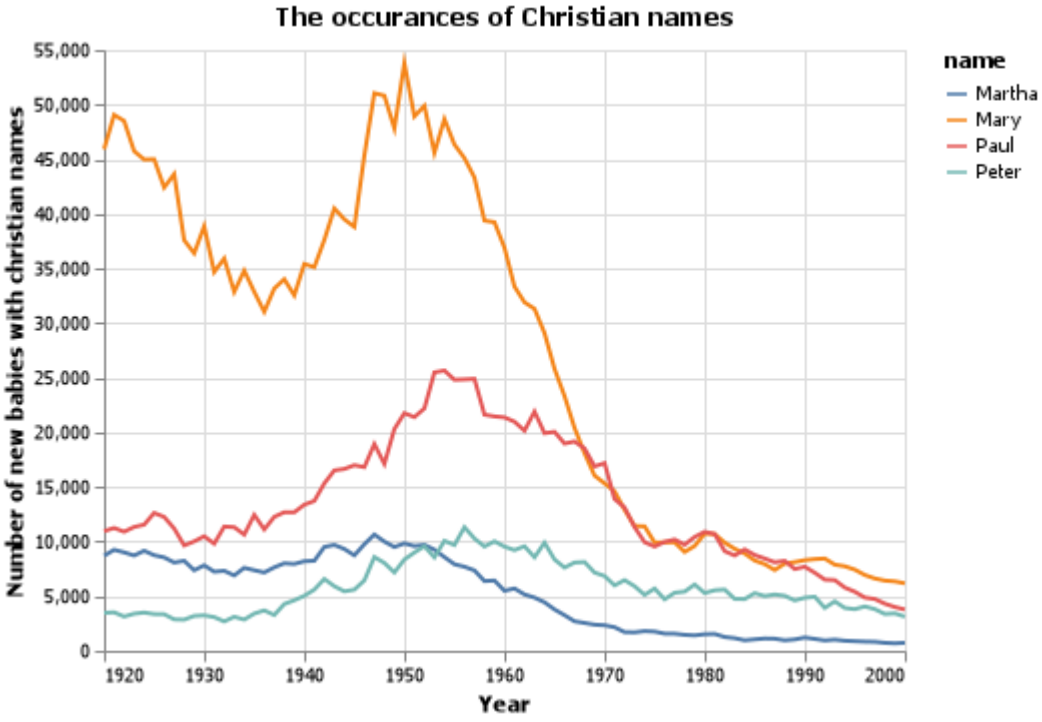
If i were to guess the age, i would say around 25 years old without looking at any data. What i would not guess would be any age over 35.



According to the data, what this shows is that the most common year of birth for Brittany is 1990. This would make the most likely age for someone with the name Brittany 31 years old. My prediction was somewhat accurate considering i was 6 years away from what the average would be.

GRAND QUESTION 3

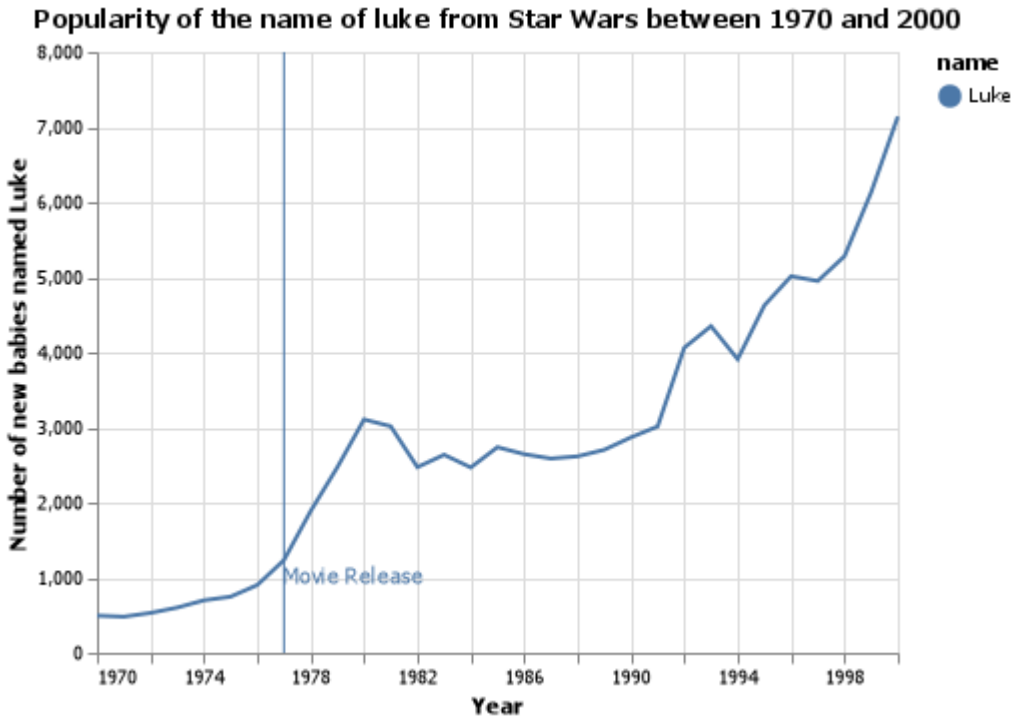
Mary, Martha, Peter, and Paul are all Christian names. From 1920 - 2000, compare the name usage of each of the four names.



From the data above we can see the change in the usage of typical christian names over time. Before the year of 1970, the name Mary was monumentally more popular than the others. However, over time the popularity drops and we see all of the names normalize at similair levels by the year 2000.

GRAND QUESTION 4

Think of a unique name from a famous movie. Plot that name and see how increases line up with the movie release.



I choose to pick Luke from Star wars to see if the release of the first movie increased the amount of new babies with the name luke. From the data we can see that before the release of star wars there is a small increase, but following star wars there is a large spike in new babies with the name luke. I think with the data it would be reasonable to say that star wars increased the amount of new babies with the name luke.

APPENDIX A (PYTHON SCRIPT)

```
# %%
#Imports data and libraries
import pandas as pd
import numpy as np
import altair as alt
url = "https://github.com/byuidatascience/data4names/raw/master/data-
raw/names_year/names_year.csv"
dat = pd.read_csv(url)

# %%
# How many unique names do we have?
pd.unique(dat.name).size
len(pd.unique(dat.name))

# %%
# How many times does my name show up?
dat.query('name == "Austin"').year.size

# %%
# Which names have been
# given the most and the least?
```

```
# https://byuidatascience.github.io/python4ds/transform.html#grouped-summaries-or-aggregations-with-.agg
# https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.groupby.html
dat_name = (dat.groupby(['name'])
    .agg(Total_all = ('Total', np.sum),
        Average_all = ("Total", np.mean))
    .reset_index())
dat_name.sort_values('Total_all').head(1).name
dat_name.sort_values('Total_all').tail(1).name

# %%
# Shows the most popular and least popular names in Utah
dat_name_state = (dat.groupby(['name'])
    .agg(Total_all = ('UT', np.sum),
        Average_all = ("UT", np.mean))
    .reset_index()
    .query("Total_all > 0")
    .sort_values("Total_all"))
print(dat_name_state.head(1).name)
dat_name_state.tail(1).name

# %%
# Graphs most popular names in utah
(alt.Chart(dat_name_state.tail(25))
    .encode(
        x = alt.X('name', sort='-y'),
        y = "Total_all")
    .mark_bar())

# %%
# Creates graph of the occurance of my name
chart = (alt.Chart(dat.query("name ==  'Austin'"),
 title = "The occurances of 'Austin")
    .encode(
        alt.X("year", axis=alt.Axis(format='.0f'), title = "Year of Birth"),
        alt.Y("Total", title="Number of Austin Births"),
        alt.Color("name")
    )
    .mark_line())
chart

# %%
# Creates data frame that shows my name historicialy over the years
dat_line = pd.DataFrame({
    "year":[1998],
    "name": ["Austin"],
    "label": ["Birth Year"],
    "y": [1000]
})
line_chart = alt.Chart(dat_line).encode(x = "year", color="name").mark_rule()
label_chart = alt.Chart(dat_line).encode(
    x = "year",
    y = "y",
    text = "label",
```

```
    color = "name").mark_text(dx=35)
chart_name = chart + label_chart + line_chart
chart_name.save("my_name_chart.png")

# %%
# Creates dataframe that shows the use of Brittany over the years
# Shows average birth year for brittany
dat_name_brittany = (dat
  .groupby(['name', 'year'])
  .sum()
  .query('name == "Brittany"')
  .reset_index()
  .filter(['name', 'year', 'Total'])
  .sort_values('Total'))

df = pd.DataFrame(dat_name_brittany)

# brittany_chart = df.plot.bar(x='year',
# title = "Instances of Brittany")
brittany_chart_graph = (alt.Chart(dat.query("name ==  'Brittany'"),
 title = "The occurances of 'Brittany")
    .encode(
        alt.X("year", axis=alt.Axis(format='.0f'), title = "Year of Birth"),
        alt.Y("Total", title="Number of brittany Births"),
        color=alt.value("#FFAA00")
    )
    .mark_line())

dat_line_brittany = pd.DataFrame({
    "year":[1996, 1986],
    "name": ["Guess", "Guess"],
    "label": ["Guess", "Guess"],
    "y": [1000, 1000]
})

line_brittany = alt.Chart(dat_line_brittany).encode(x = "year",
color="name").mark_rule()
brittany_chart = brittany_chart_graph + line_brittany

# %%
# Saves chart for brittany
brittany_chart.save("brittany_chart.png")

# %%
# creates data frame for christian names
chart_cr = dat.query("name in ['Mary', 'Martha', 'Peter', 'Paul'] & (year >= 1920
& year <= 2000)")

# %%
# creates and saves christian chart
(alt.Chart(chart_cr,
 title = "The occurances of Christian names")
    .encode(
        alt.X('year:Q', axis=alt.Axis(format='.0f'), title = "Year"),
```

```
        alt.Y("Total", title="Number of new babies with christian names"),
        alt.Color("name"),
    )
    .mark_line().save("christian_chart.png"))

# %%
# creates chart for luke
arnold_chart = (alt.Chart(dat.query("name ==  'Luke' & (year >= 1970 & year <=
2000)"),
 title = "Popularity of the name luke from Star Wars between 1970 and 2000")
    .encode(
        alt.X("year", axis=alt.Axis(format='.0f'), title = "Year"),
        alt.Y("Total", title="Number of new babies named Luke"),
        alt.Color("name")
    )
    .mark_line())

# %%
# creates dataframe and combines chart data
dat_line_arnold = pd.DataFrame({
    "year":[1977],
    "name": ["Luke"],
    "label": ["First Star Wars Movie Release"],
    "y": [1000]
})
line_chart_arnold = alt.Chart(dat_line_arnold).encode(x = "year",
color="name").mark_rule()
label_chart_arnold = alt.Chart(dat_line_arnold).encode(
    x = "year",
    y = "y",
    text = "label",
    color = "name").mark_text(dx=35)
chart_arnold = arnold_chart + label_chart_arnold + line_chart_arnold

# %%
# Saves chart for arnold
chart_arnold.save("arnold_chart.png")

# %%
```