# NAME DATA, USA

HTTPS://WWW.KAGGLE.COM/DATAGOV/USA-NAMES

Krystina Moses, DSC 530

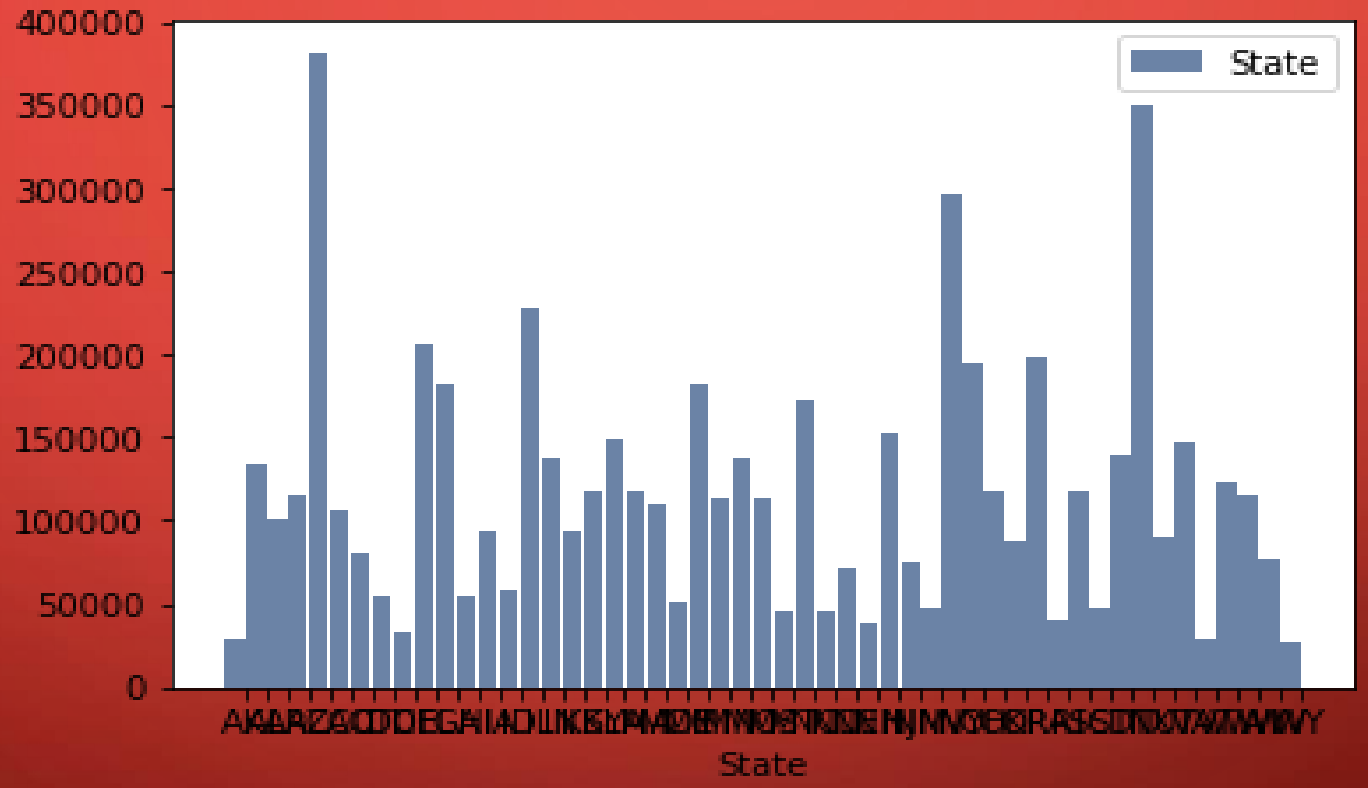
# STATISTICAL QUESTIONS

- What are the most common male and female names throughout the years?
- Which year had more males vs females?
- Over all the years, are there more males or females?
- Which states have more males? Females?
- What year had the most name occurrences?

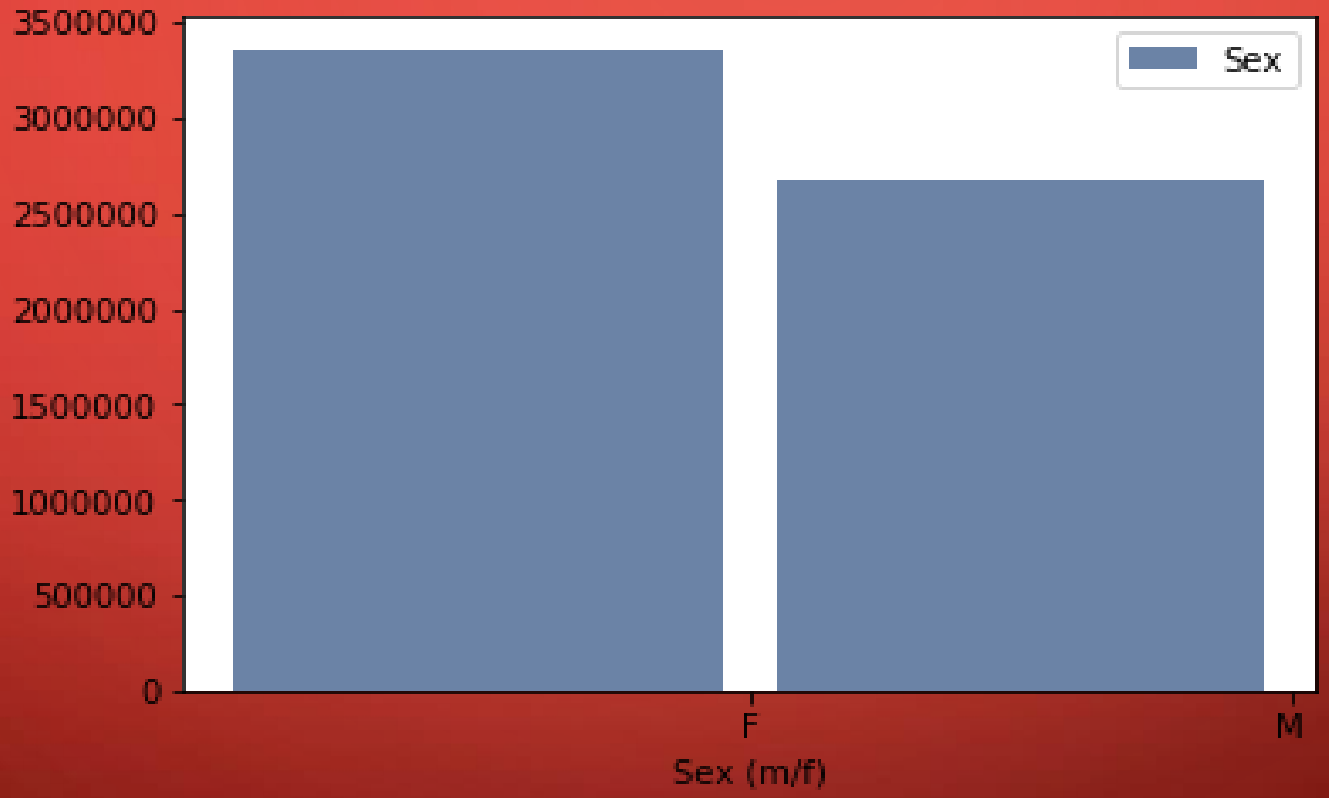
# VARIABLES: MEANING

- STATE: 2-digit state code
- GENDER: Sex (M = Male, F = Female)
- YEAR: 4-digit year of birth
- NAME: Name at birth
- NUMBER: Number of occurrences of the name

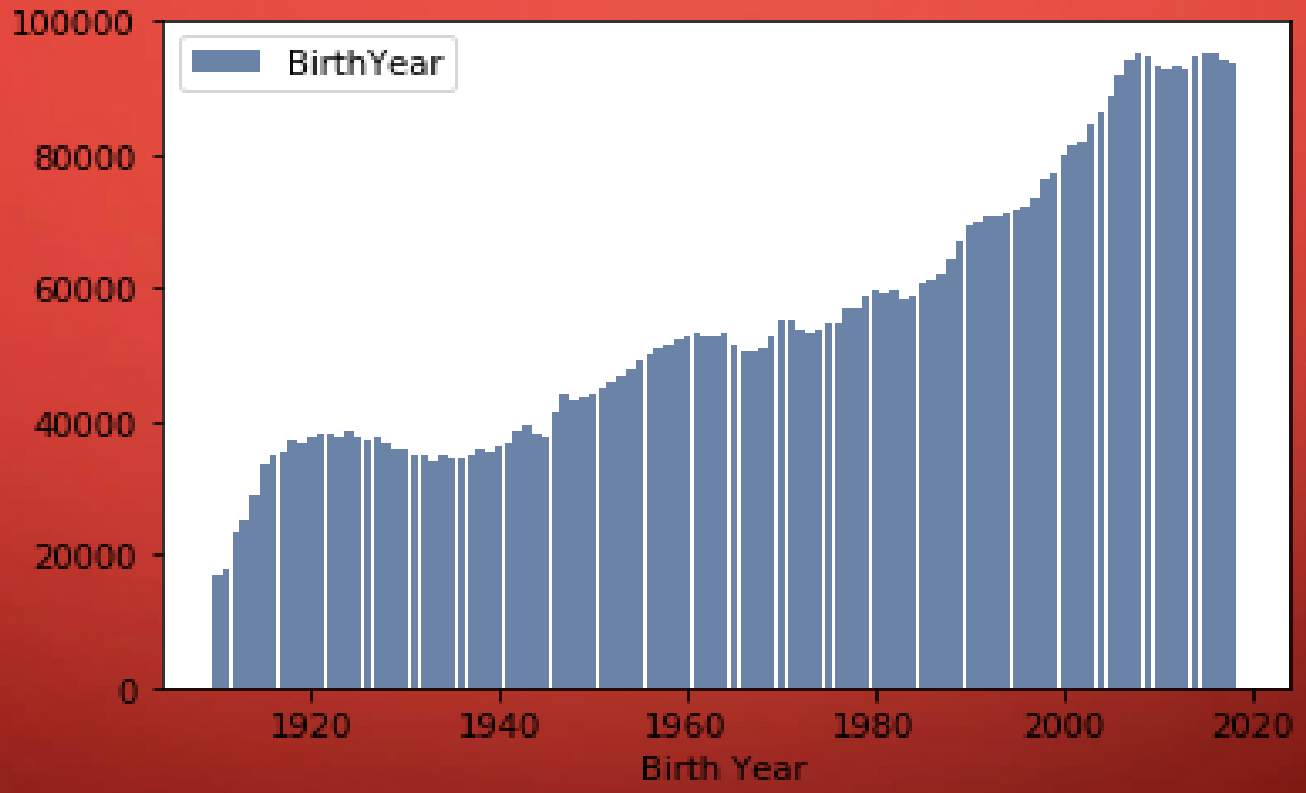
# HISTOGRAM OF STATE VARIABLE



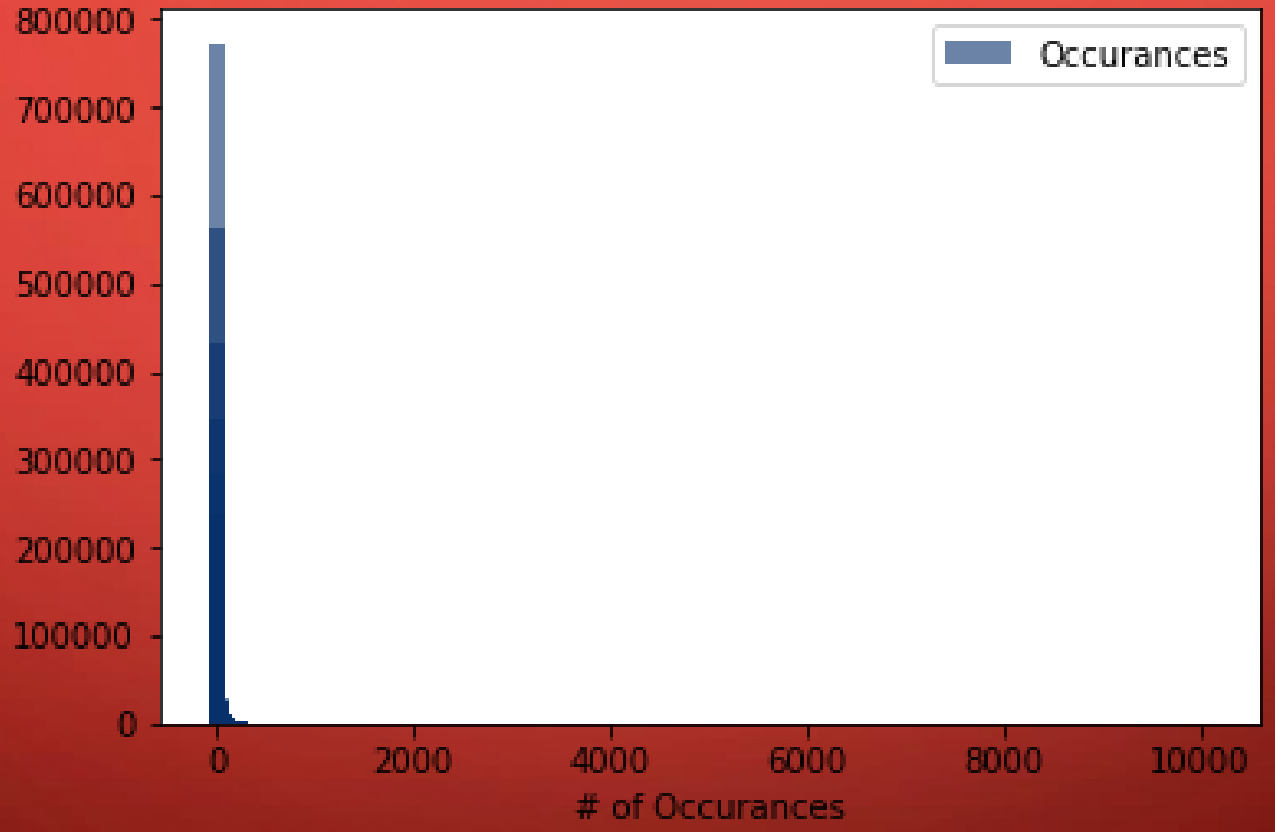
# HISTOGRAM OF SEX (M/F) VARIABLE



# HISTOGRAM OF BIRTH YEAR VARIABLE



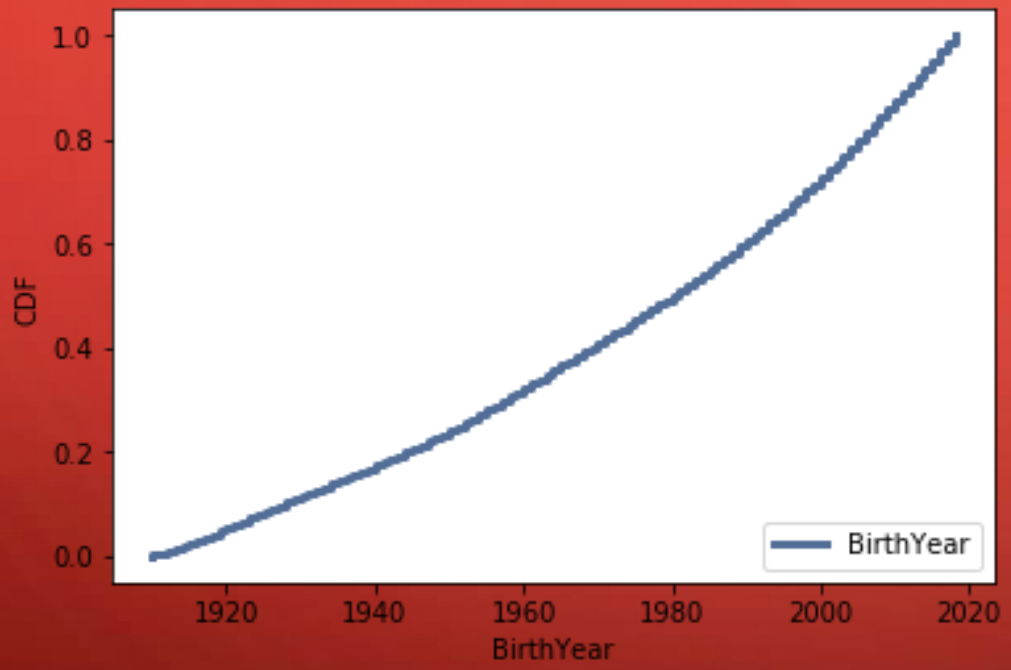
# HISTOGRAM OF OCCURRENCES VARIABLE



# VARIABLE CHARACTERISTICS

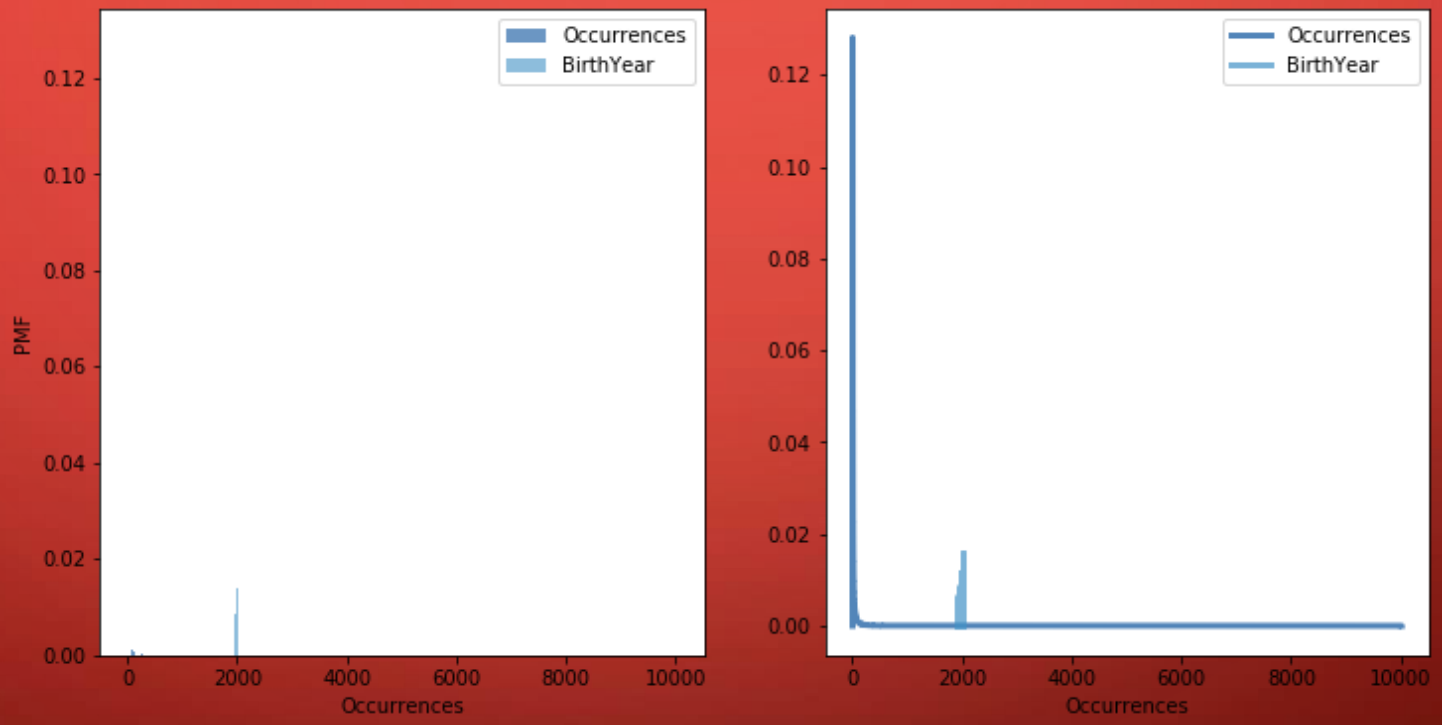
- Birth Year –
  - Mean: 1975.2
  - Variance: 934.3
  - Standard Deviation: 30.6
- Occurrences –
  - Mean: 51.6
  - Variance: 31097.6
  - Standard Deviation: 176.3

# CUMULATIVE DISTRIBUTION FUNCTION (CDF)

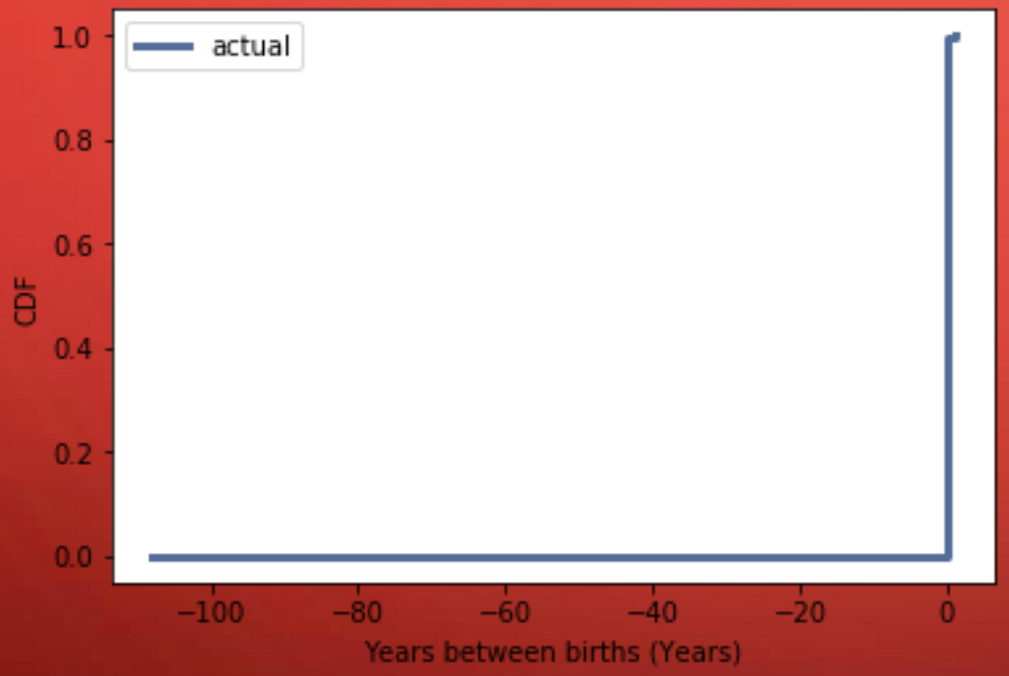



- CDF range from .002 - 0.98 over the years.
- The earlier years are in the .02 percentile and the later years are closer to the 98 percentile.

# COMPARE TWO SCENERIOS USING A PMF

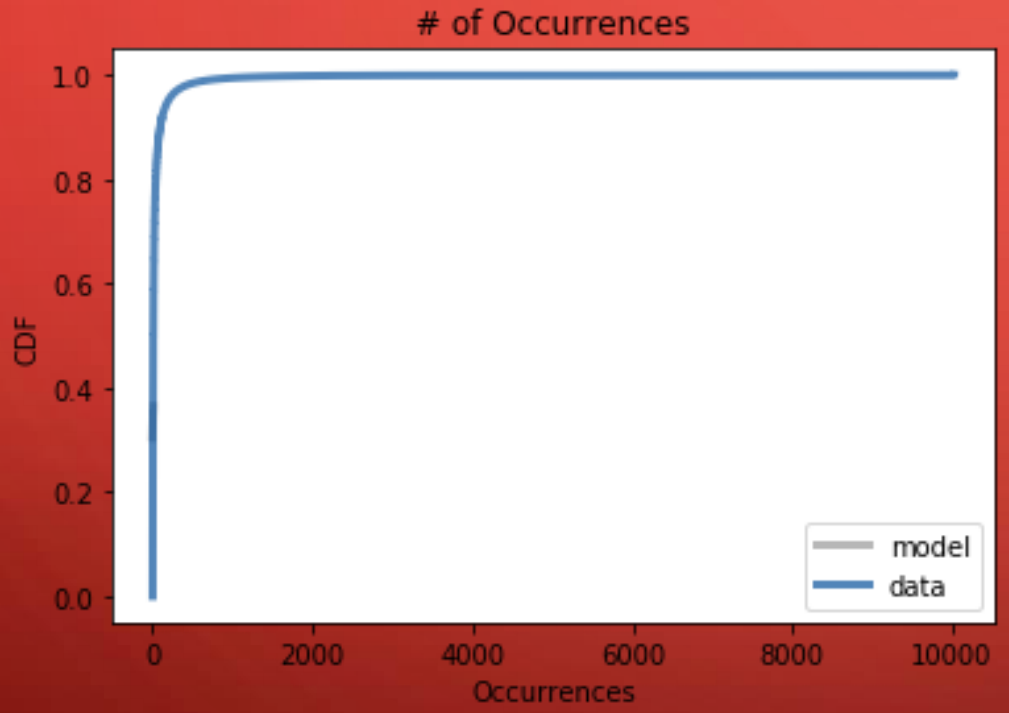


# EXPONENTIAL DISTRIBUTION




- Exponential distribution is not being shown for the Years between births

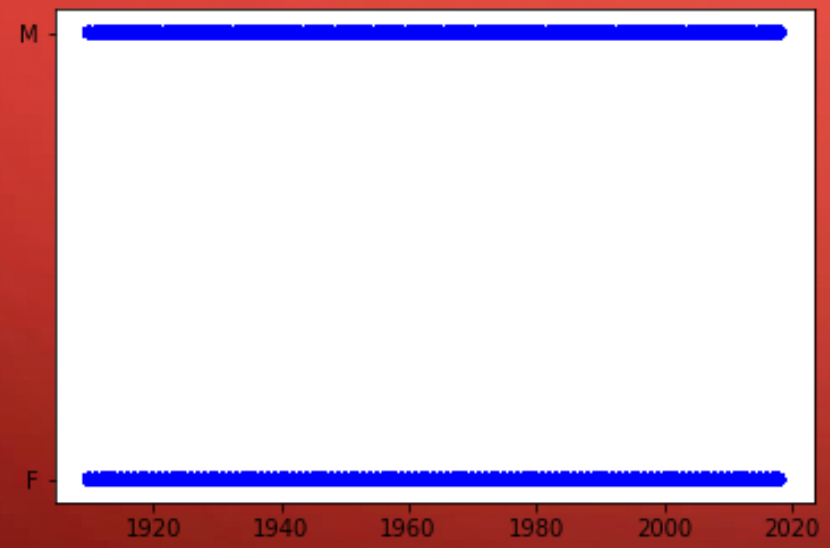
# NORMAL DISTRIBUTION



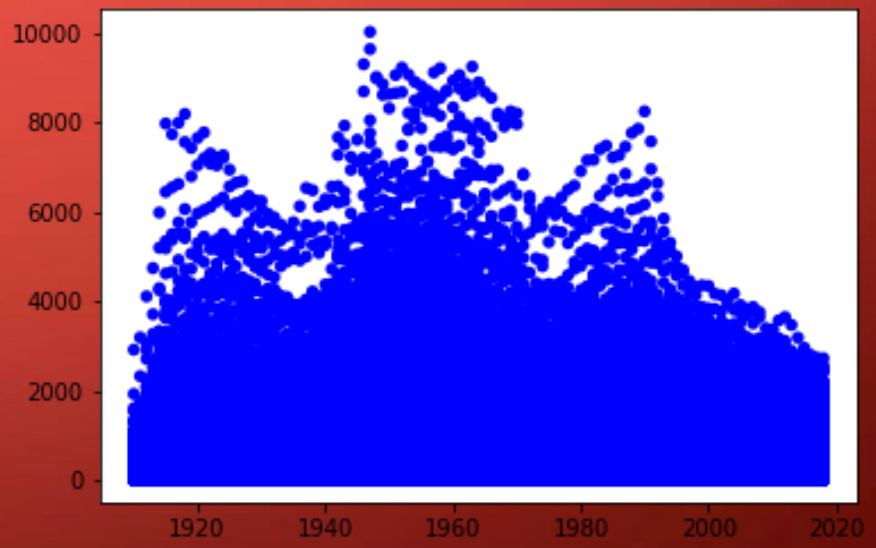

- Normal distribution plot for # of occurrences of names

# SCATTER PLOTS COMPARING TWO VARIABLES

## BIRTH YEAR VERSUS SEX




## BIRTH YEAR VERSUS OCCURRENCE




Pearson's : -.001

Spearman's: 0.0001

# HYPOTHESIS TESTING

- It was determined that there was no relationship between the gender of the baby and how many times either male or female had occurred within the year.

# REGRESSION ANALYSIS

- OLS Regression Results
- Dep. Variable: nsdf2.BirthYear R-squared: 0.000
- Model: OLS Adj. R-squared: 0.000
- Method: Least Squares F-statistic: 2.034
- Date: Sat, 10 Aug 2019 Prob (F-statistic): 0.154
- Time: 12:46:32 Log-Likelihood: -1.3646e+05
- No. Observations: 188016 AIC: 2.729e+05
- Df Residuals: 188014 BIC: 2.729e+05
- Df Model: 1
- Covariance Type: nonrobust
- coef std err t P>|t| [0.025 0.975]
- Intercept 2017.4990 0.001 1.63e+06 0.000 2017.497 2017.501
- nsdf2.Occurrences -2.049e-05 1.44e-05 -1.426 0.154 -4.87e-05 7.67e-06
- Omnibus: 640545.310 Durbin-Watson: 0.004
- Prob(Omnibus): 0.000 Jarque-Bera (JB): 31334.646
- Skew: 0.007 Prob(JB): 0.00
- Kurtosis: 1.000 Cond. No. 92.7