Exploring Bike Segments in Idaho

Leo Bomboy, Nathan Nguyen, Andrew Plum, Noah Rieth, Anna Ronaye, Jonna Waage

Why We Chose the Strava Dataset

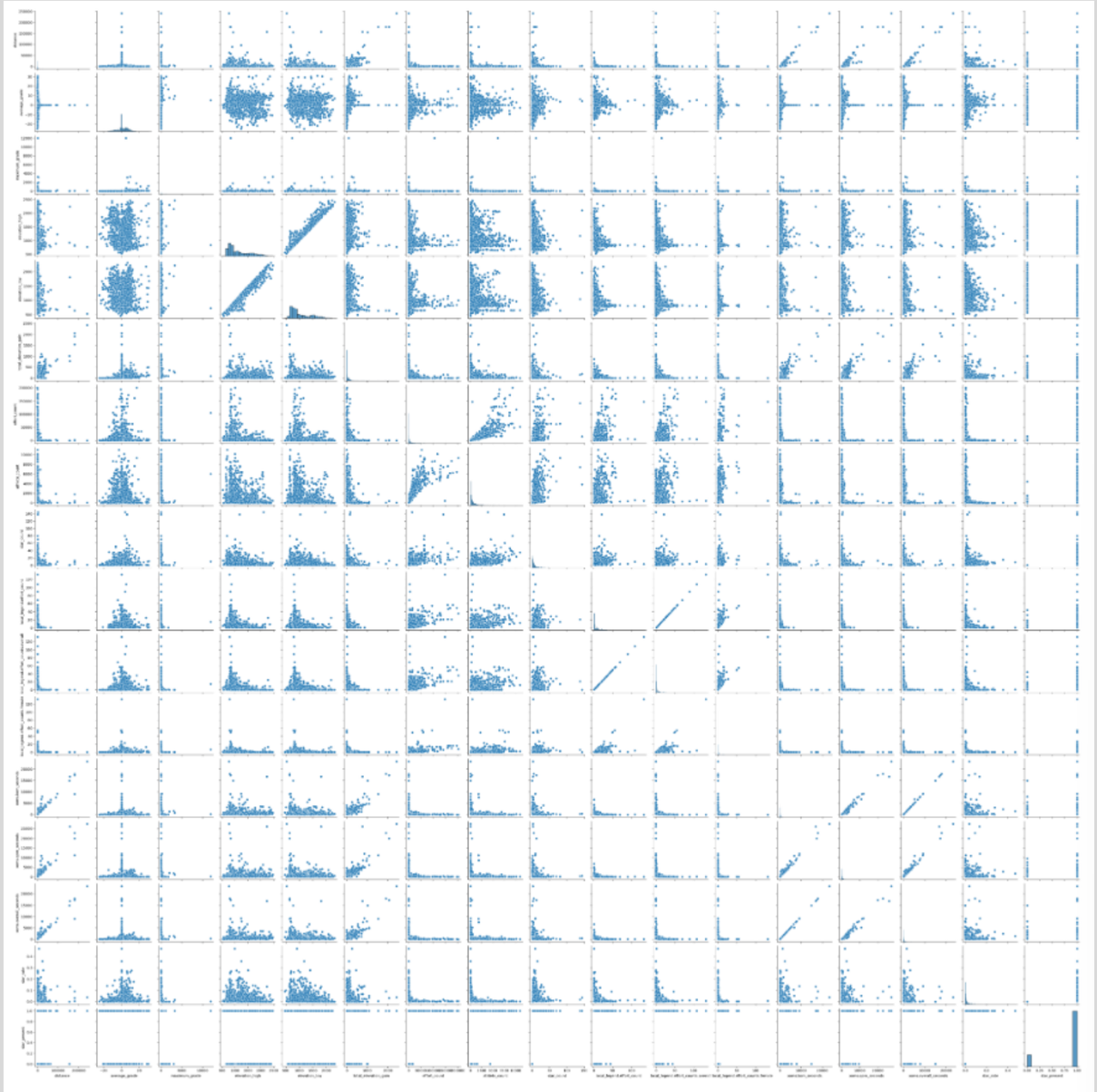
- Goals
  - Learn what factors correlate with an athlete's choice of path
  - Investigate ride speed among athletes of varying experience/fitness
- Reasons:
  - Richness of Data
  - Availability and Accessibility
  - Diverse Data Source
- How the Data was Found and Managed:
  - Pulled via the Strava API
  - Relevant data was extracted and organized based on the research goals

Methods of Discovery

- Data Formats Used:
  - JSON converted to CSV
- Metadata Standards/Conventions Applied:
  - Dublin Core Metadata Elements to organize data
  - Standardize Metadata enhanced data description
- Method(s) of Discovery and Access:
  - Python API used to pull Strava data
  - Dublin Core Metadata Elements were created as the metadata standard for this dataset
- Impact on Process:
  - Helped in organizing data and metadata
  - Hindered by the need to decide which data fields to focus for analysis
  - Overall facilitated data comprehension and analysis within the group

Mind Map Representation of Data Relatedness

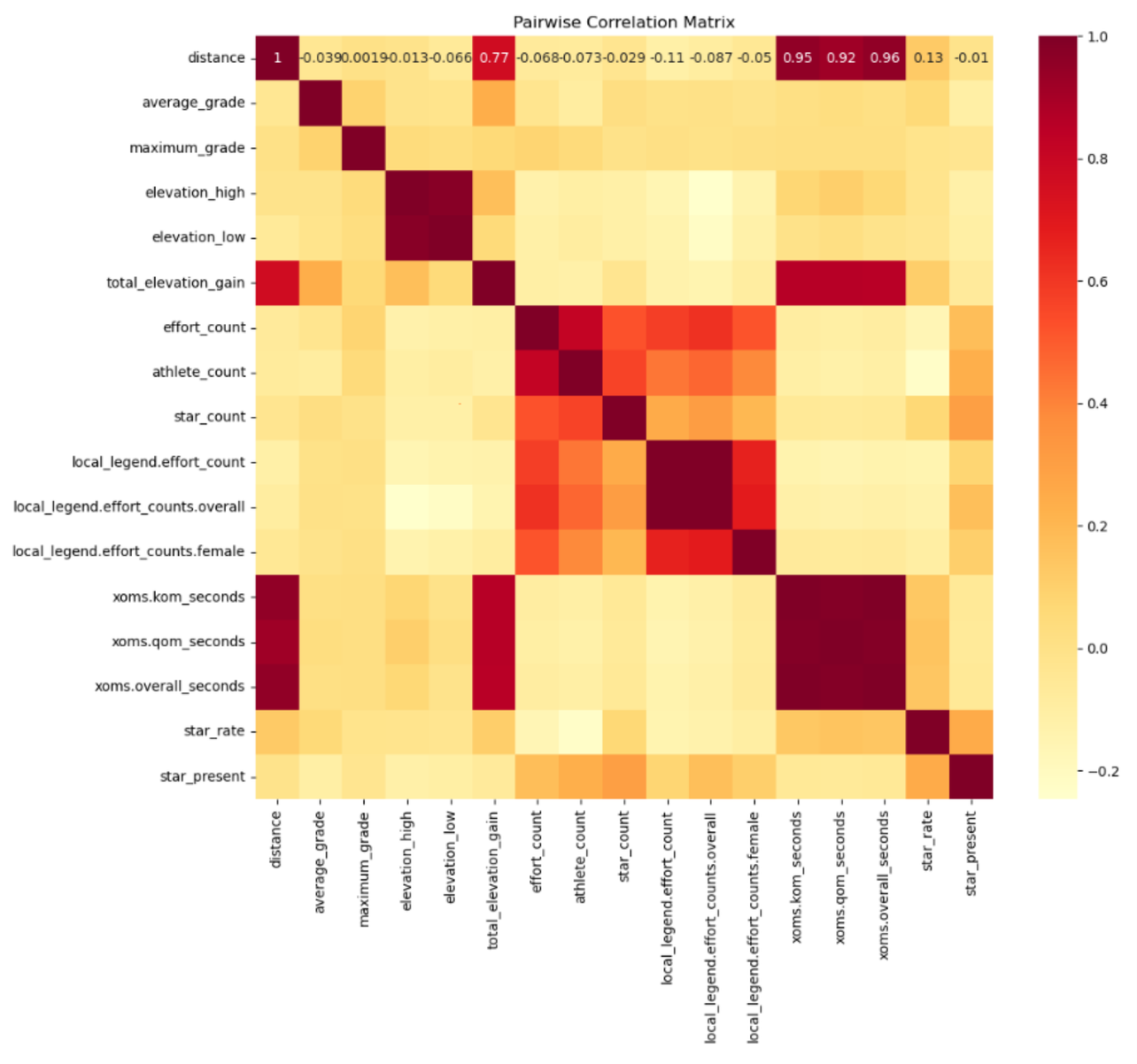
Initial Exploratory Visualization (1)



What trends exist in the data?

What questions can we ask?

Initial Exploratory Visualization (2)



What trends exist in the data?

What questions can we ask?

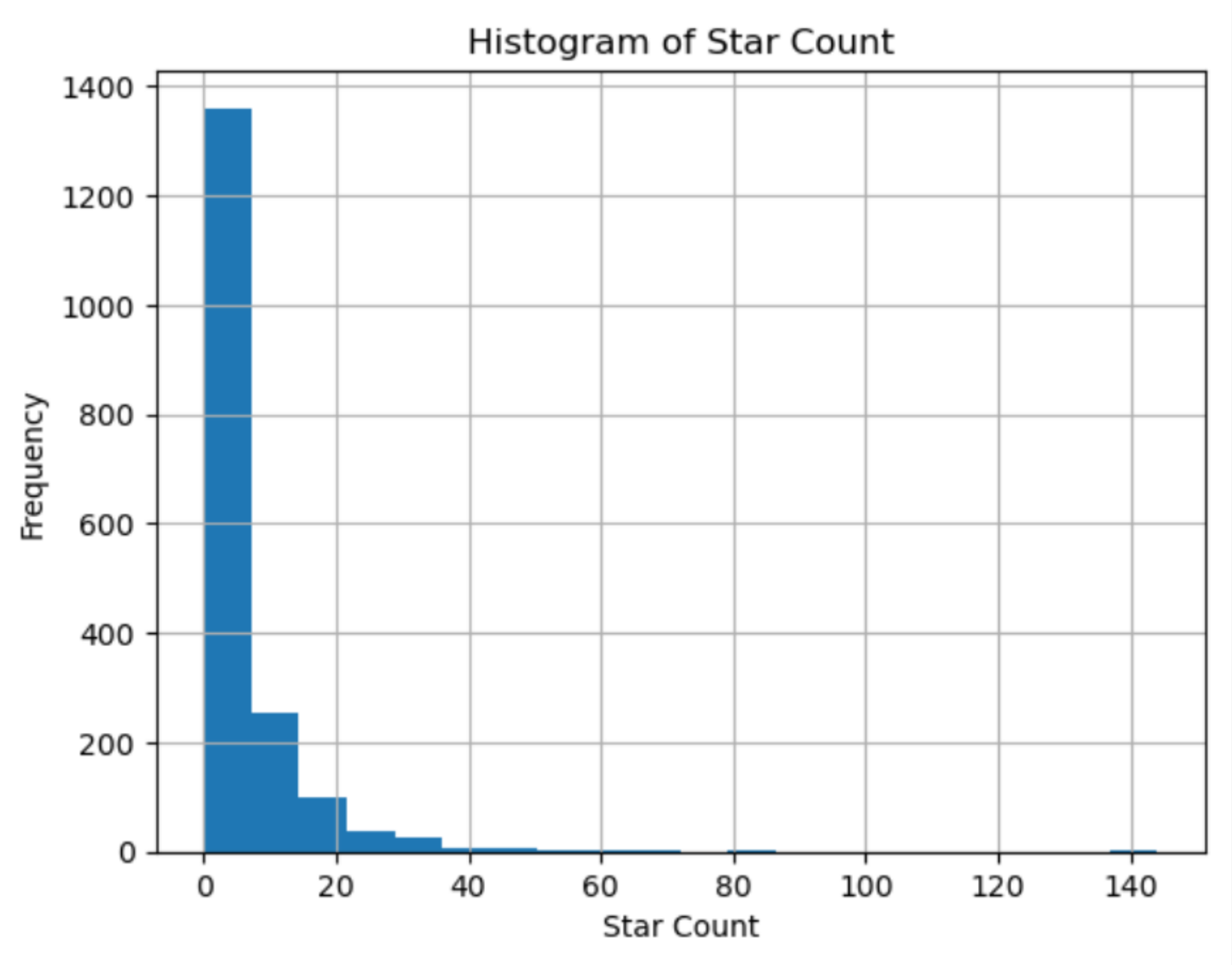
Questions Related to Goal

- Investigation Aim:
  - Explore what factors correlate with an athlete's choice of paths
    - Star count is a the data attribute that indicates how favored a trail is by athletes
  - Investigate ride speeds among athletes of varying experiences/fitness levels
- Hypotheses:
  - The choice of paths taken by athletes are correlated with factors such as distance, average grade, maximum grade, elevation changes, and effort count
  - Path Characteristics (e.g elevation, age) will significantly influence ride speed irrespective of rider characteristics
- Initial Analysis Approach:
  - Regression Analysis

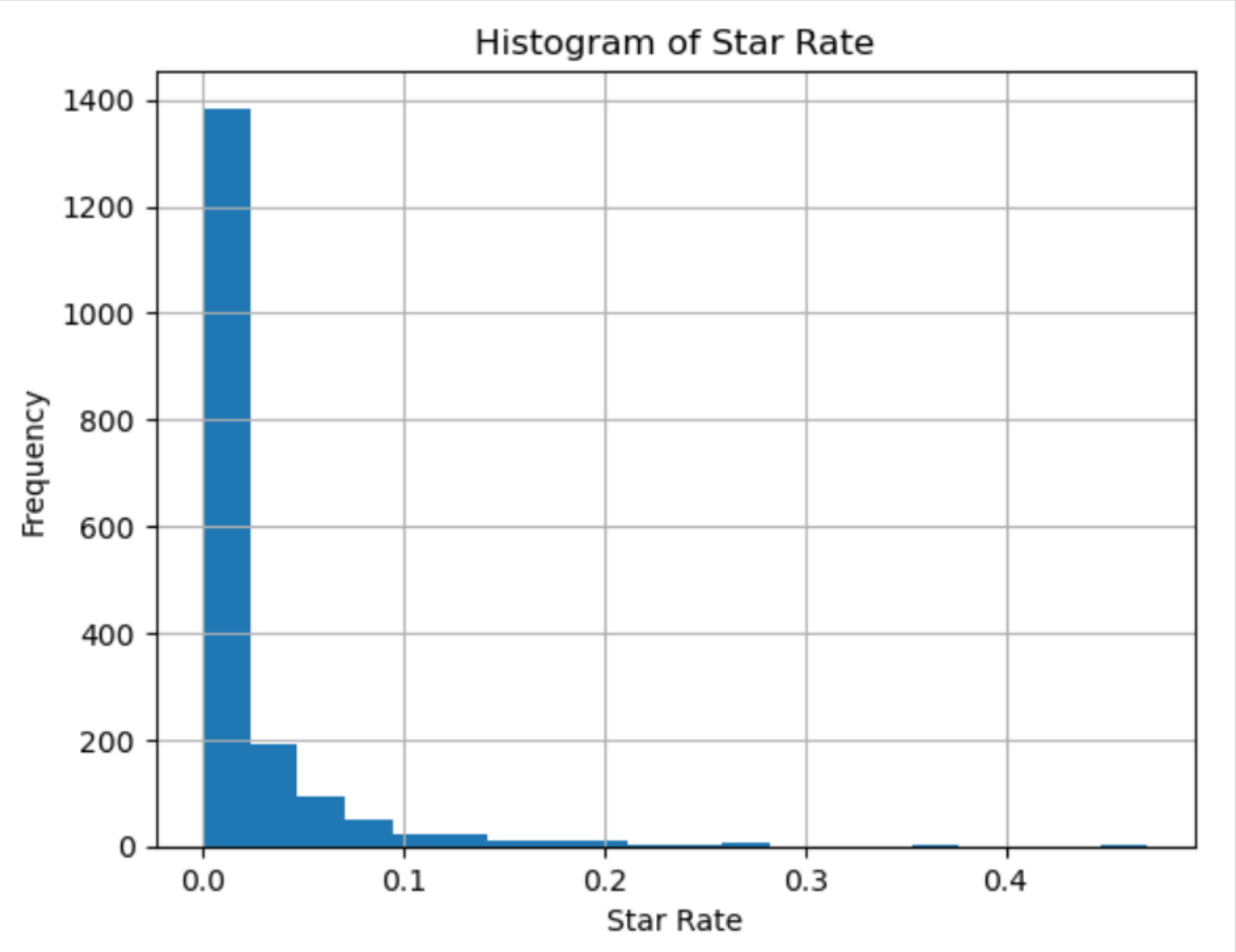
Attack Plan of First Hypothesis

- First Hypothesis:
  - Whether a trail is starred by athletes is correlated with factors such as distance, average grade, maximum grade, elevation changes, and effort count.
- The API did not have data of whether an individual athlete starred a path
  - This means it needs to be derived
    - We did have a star count and an athlete count
      - So attributes star rate (star count / athlete count) and star present (0 if no stars for trail and 1 if trail has at least 1 star) were derived
- The question could now be explored as a binary classification problem
  - This meant we could use logistic regression, KNN, SVM, binary neural network classifiers as our models

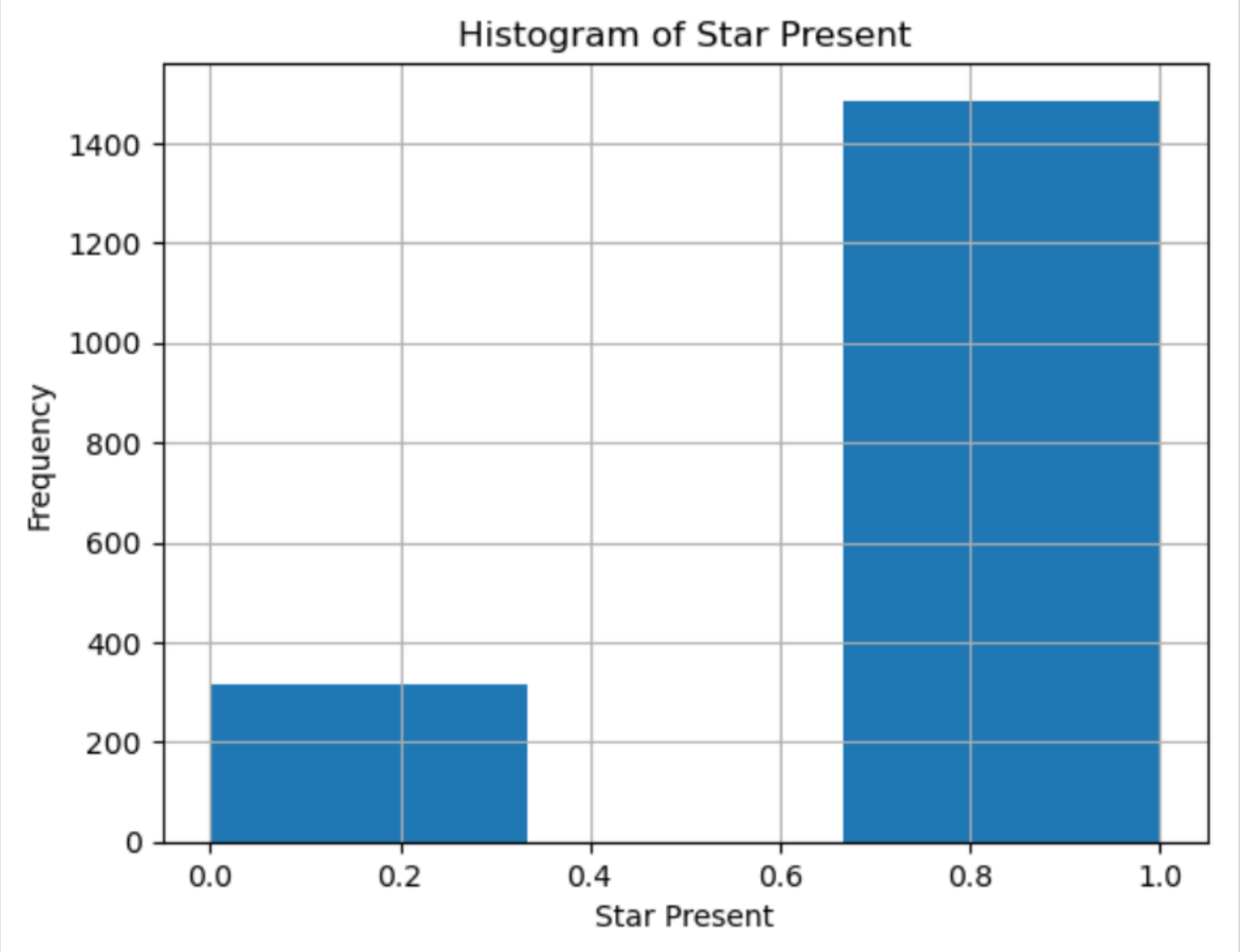
First Hypothesis Related Visualization (1)



First Hypothesis Related Visualization (2)



First Hypothesis Related Visualization (3)



Logistic Regression Results

```
Optimization terminated successfully.
         Current function value: 0.374527
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:           star_present   No. Observations:                 1799
Model:                         Logit   Df Residuals:                     1791
Method:                          MLE   Df Model:                            7
Date:                Sat, 30 Mar 2024   Pseudo R-squ.:                 0.1911
Time:                        17:45:32   Log-Likelihood:                -673.77
converged:                       True   LL-Null:                       -832.96
Covariance Type:            nonrobust   LLR p-value:                 7.159e-65
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept          0.9534      0.208      4.579      0.000       0.545       1.362
distance         3.256e-06    9.6e-06      0.339      0.735   -1.56e-05    2.21e-05
average_grade     -0.0284      0.012     -2.454      0.014      -0.051      -0.006
maximum_grade     -0.0007      0.000     -1.517      0.129      -0.001       0.000
elevation_high    -0.0010      0.001     -1.031      0.303      -0.003       0.001
elevation_low      0.0008      0.001      0.791      0.429      -0.001       0.003
total_elevation_gain  0.0007      0.001      0.743      0.457      -0.001       0.002
effort_count       0.0005   6.08e-05      8.783      0.000       0.000       0.001
==============================================================================

Possibly complete quasi-separation: A fraction 0.13 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

First Logit Model

```
Optimization terminated successfully.
         Current function value: 0.377070
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:           star_present   No. Observations:                 1799
Model:                         Logit   Df Residuals:                     1796
Method:                          MLE   Df Model:                            2
Date:                Sat, 30 Mar 2024   Pseudo R-squ.:                 0.1856
Time:                        18:44:47   Log-Likelihood:                -678.35
converged:                       True   LL-Null:                       -832.96
Covariance Type:            nonrobust   LLR p-value:                 7.107e-68
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept          0.6970      0.086      8.087      0.000       0.528       0.866
average_grade     -0.0291      0.010     -2.908      0.004      -0.049      -0.009
effort_count       0.0005   5.97e-05      9.023      0.000       0.000       0.001
==============================================================================

Possibly complete quasi-separation: A fraction 0.13 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

Second Logit Model

Going Forward

- Utilize binary classification models alongside.
  - K-nearest-neighbor
  - Support vector machines
  - Binary neural network classifier

Overall Data Management Plan

- Creation of Logical Collections
  - Two collections based on the two hypothesized questions will exist. With two databases, analysis can be easier due to split research goals.
- Physical Data Handling
  - Raw data as well as the two databases (CSVs) will be stored in the Google Drive Cloud Environment. Allowing for enhanced security, availability, and easier custodianship.

Overall Data Management Plan Cont.

- Interoperability Support
  - Data collected from the Strava API will be thoroughly investigated to ensure compatibility with our open source analysis tools. Which will mostly consist of Google Drive, Visual Studio, and other open source office software programs.
- Security Support
  - Robust measures such as cloud storage, in transit/at rest encryption, and access controls will be implemented. This allows our data to have sufficient confidentiality, integrity, and security.

Overall Data Management Plan Cont.

- Data Ownership
  - Responsibility for the data will be based upon roles assigned to group members. To ensure a chain of custody going forwards.
- Metadata Collection, Management, and Access
  - Collected metadata consisting of rider and path information will be stored in the cloud. This will also allow authorized individuals and stakeholders to access data organized according to the Dublin Core for Metadata Standards.

Overall Data Management Plan Cont.

- Persistence
  - Google Drive allows for high persistence, with our databases existing in several high availability locations at once.
- Discovery
  - As data is ingested, it will be thoroughly integrated into the metadata standards mentioned. Categorization allows for quicker analysis and referencing.

Overall Data Management Plan Cont.

- Data Dissemination and Publication
  - During the project, distribution of the dataset will only occur within the team and with our supervisors. After which data will be uploaded to an open source research base such as Github.