

JongWook Choe

San Francisco, CA, 94102 | (530) 760-6057

1994wook@gmail.com | <https://linkedin.com/in/1994wook> | <https://wook22.github.io/PORTFOLIO/>

Summary

Data analyst with expertise in cleaning, transforming and analyzing large datasets using advanced statistical methods and machine learning techniques. Skilled in data visualization tools such as Tableau and proficient in programming languages R, Python, and SQL. Comfortable communicating complex data-driven insights to both technical and non-technical stakeholders using strong communication skills and creating compelling data visualizations.

Technical Skills

- Python (Pandas, Numpy, Seaborn, Hadoop, Spark)
- Web Scraping (Beautifulsoup, Selenium), APIs
- SQL (Postgres, Bigquery)
- Tableau
- Javascript (D3.js, Plotly, Leaflet)
- HTML
- Machine Learning (Supervised, Unsupervised)

Experience

Summer Institute Fellow

Jun. 2022 - Oct. 2022

California Policy Lab

Berkeley, CA

- Collaborated with a team to construct an ETL (extract, transform, load) data pipeline by merging multiple data frames of nationwide homelessness datasets.
- Analyzed and examined the trend of homelessness in California over the past twenty years and identified the factors contributing to the fluctuations in the homeless population.
- Managed data preparation process of a longitudinal dataset from the Department of Housing and Urban Development dating back to 2015, with Continuum of Care as a primary variable consisting of 100,000+ rows.
- Developed and presented findings to a multidisciplinary team, providing important policy implications for addressing homelessness in California and other regions facing similar challenges.

Public Data Intern

Jun. 2021 - Aug. 2021

National Information Society Agency, NIA

Seoul, South Korea

- Worked as a team to support the government's "Open Data Activation Policy" to promote the use of public data.
- Optimized the data maintenance tool by automating the data inspection process with Python and achieved higher data accuracy from 91% to 99%.
- Provided data maintenance support for the public data portal, www.data.go.kr

Marketing Intern

Jul. 2018 ~ Sep. 2018

Daily Beer, Inc.

Seoul, South Korea

- Created statistical reports on new delivery businesses to provide insights into commercial analysis and market research.
- Organized and shaped social networking business campaigns based on research.

Projects

Vacation Destination Hotel Search Dashboard

Aug. 2023

[https://github.com/Wook22/Hotel_California_Dashboard]

- Examined vacation destinations across California based on hotels, their ratings, and pricing using Priceline APIs.
- Leveraged API ensuring data accuracy created hotel price data in a JSON format and securely stored within a MongoDB database. Converted hotel pricing data into JSON format for secure storage in a MongoDB database.
- Utilized Flask for data retrieval and integrated two distinct HTML files for a user-friendly interface.
- Presented hotel details such as names, addresses, scores, and rating distribution in a dynamic table and chart.
- Provided insight to help people make informed decisions about hotel selections and vacation plans.

Institution-level College Scorecard Analysis

May. 2023

[https://github.com/Wook22/Education_Institution_Data_Analysis]

- Examined U.S. Department of Education data to compare public and private colleges and gain insights into the

similarities and differences between institution types.

- Extracted and transformed socioeconomic data into five locations: ethnicity, major, type, and financial.
- Identified a higher prevalence of open admission colleges and conducted hypothesis testing to establish the statistical difference in admission costs between public and private schools,
- Provided affordability and accessibility in higher education for people's decision-making and policy development.

Consideration of Pipe Service Life

Mar. 2023

[https://github.com/Wook22/Pipelife_Analysis]

- Discovered the main factors that contributed to breakages of the pipe and predicted potential damage.
- Cleaned and preprocessed actual pipes and past breakage information and merged datasets based on a native pipe ID as the primary key variables.
- Conducted data cleaning and preprocessing on client data and merged datasets based on a native pipe ID as a primary key variable.
- Identified material and length as the factors in pipe failure through visualization and logistic regression.
- Provided insight into factor contributions to pipe breakages and prediction models for identifying pipe failures.

MLB All-Star Game Prediction

Jun. 2022

[https://github.com/Wook22/MLB_AllStar_Analysis]

- Forecasted the 2022 MLB All-Star Game player roster using 2021 season player stats and biometric data.
- Acquired data through web scraping utilizing the Python library, BeautifulSoup.
- Visualized player biometric data to identify relevant patterns and trends for All-Star Game selection.
- Identified key factors such as batting average and age employed logistic regression and generated a player list based on the players' statistics of the first half 2022 season.

NASA Asteroid Classification

May. 2022

[https://github.com/Wook22/Asteroids_Classification_Analysis]

- Examined the sizeable public dataset of asteroid characteristics and historical collisions from NASA's database to assess high-risk asteroids colliding with Earth.
- Oversampled the skewed dataset using the SMOTE algorithm to balance the data and achieve higher accuracy before model selection.
- Predicted high-risk asteroid collisions with greater accuracy using the decision tree algorithm compared to other employed models, enabling the development of more effective preventative strategies,

Data Analysis on Prevalence of Byssinosis

Mar. 2022

[https://github.com/Wook22/Byssinosis_Data_Analysis]

- Conducted research on the prevalence of byssinosis in a North Carolina cotton textile company, and which factors contributed to its occurrence.
- Formulated linear regression and factor-level interaction plots to observe trends and identify high-risk groups.
- Used stepwise forward regression with AIC to find the best-fit model to find the most relevant variables, and then tested with LR tests and other statistical tools.
- Generated best-fitted model to uncover meaningful association in the data, concluding that smoking workers who have been exposed to dusty environments for over 20 years have the highest risk of byssinosis.

Certificates

- **Google Analytics** Individual Qualification
- **Google Data Analytics Specialization** certificate programs
- **Stanford Machine Learning** certificate programs
- **SQL for Data Science** certificate programs

Education

Berkeley Data Analytics Boot Camp, Berkeley, CA, USA

Oct. 2023

- Advanced understanding of Excel, Python, and R programming, JavaScript charting, SQL database, Tableau, and machine learning.

University of California Davis, Davis, CA, USA

Jun. 2022

- Bachelor of Science in Statistics; minor in Mathematics