

# **Introduction to Data Science**

**STAT 3255/5255 @ UConn, Fall 2024**

Jun Yan and Students in Fall 2024:

Zachary Blanchard, Sara Clokey      Julia Mazzola

2024-10-07



# Table of contents

<b>Preliminaries</b>	<b>1</b>
Sources at GitHub . . . . .	1
Midterm Project . . . . .	2
Final Project . . . . .	2
Adapting to Rapid Skill Acquisition . . . . .	2
Wishlist . . . . .	3
Presentation Orders . . . . .	7
Course Logistics . . . . .	8
Presentation Task Board . . . . .	8
Final Project Presentation Schedule . . . . .	11
Contributing to the Class Notes . . . . .	11
Homework Requirements . . . . .	12
Practical Tips . . . . .	13
Data analysis . . . . .	13
Presentation . . . . .	13
My Presentation Topic (Template) . . . . .	13
Introduction . . . . .	13
Sub Topic 1 . . . . .	13
Sub Topic 2 . . . . .	14
Sub Topic 3 . . . . .	14
Conclusion . . . . .	14
<b>1 Introduction</b>	<b>15</b>
1.1 What Is Data Science? . . . . .	15
1.2 Expectations from This Course . . . . .	16

## Table of contents

1.3	Computing Environment . . . . .	17
1.3.1	Command Line Interface . . . . .	17
1.3.2	Python . . . . .	18
<b>2</b>	<b>Project Management</b>	<b>21</b>
2.1	Set Up Git/GitHub . . . . .	21
2.2	Most Frequently Used Git Commands . . . . .	22
2.3	Tips on using Git: . . . . .	23
2.4	Pull Request . . . . .	23
<b>3</b>	<b>Reproducible Data Science</b>	<b>25</b>
3.1	Introduction to Quarto . . . . .	25
3.2	Compiling the Classnotes . . . . .	26
3.2.1	Set up your Python Virtual Environment . . . . .	26
3.2.2	Clone the Repository . . . . .	27
3.2.3	Render the Classnotes . . . . .	28
<b>4</b>	<b>Python Refreshment</b>	<b>29</b>
4.1	Know Your Computer . . . . .	29
4.1.1	Operating System . . . . .	29
4.1.2	File System . . . . .	30
4.2	The Python World . . . . .	31
4.3	Standard Library . . . . .	31
4.4	Important Libraries . . . . .	33
4.5	Writing a Function . . . . .	34
4.5.1	Monte Hall . . . . .	35
4.6	Variables versus Objects . . . . .	36
4.7	Number Representation . . . . .	39
4.7.1	Integers . . . . .	39
4.7.2	Floating Number . . . . .	40
4.8	Virtual Environment . . . . .	42
<b>5</b>	<b>Communication in Data Science</b>	<b>45</b>
5.1	Introduction . . . . .	45

## Table of contents

5.2	Importance & Application of Communication . . . . .	45
5.3	General Communication Skills . . . . .	46
5.3.1	Verbal Communication Skills . . . . .	46
5.3.2	Non-Verbal Communication Skills . . . . .	47
5.3.3	Visual Communication Skills . . . . .	47
5.3.4	Written Communication Skills . . . . .	47
5.4	Communication in Data Science . . . . .	48
5.4.1	Identify your Audience . . . . .	48
5.4.2	Utilize Data Visualization . . . . .	48
5.4.3	Focus on Data Communication Skills . . . . .	49
5.4.4	Give Space for Questions and Feedback . . . . .	49
5.5	Further Learning . . . . .	50
<b>6</b>	<b>Data Manipulation</b>	<b>51</b>
6.1	Introduction . . . . .	51
6.2	Import/Export Data . . . . .	52
6.2.1	Subsection 1 . . . . .	52
6.2.2	Subsection 2 . . . . .	52
6.2.3	Extended Reading . . . . .	52
6.3	SQL . . . . .	52
6.3.1	Subsection 1 . . . . .	52
6.3.2	Subsection 2 . . . . .	52
6.3.3	Extended Reading . . . . .	52
6.4	NYC Crash Data . . . . .	53
6.5	Cross-platform Data Format <b>Arrow</b> . . . . .	56
6.6	Using the Census Data . . . . .	58
<b>7</b>	<b>Visualization</b>	<b>63</b>
7.1	Data Visualization with Plotnine . . . . .	63
7.1.1	Introduction . . . . .	63
7.1.2	What is Plotnine? . . . . .	63
7.1.3	Installing Plotnine . . . . .	64
7.1.4	Scatterplot . . . . .	65
7.1.5	Bar Chart . . . . .	82

## Table of contents

7.1.6	Histogram . . . . .	89
7.1.7	Line Chart . . . . .	93
7.1.8	Conclusion . . . . .	99
7.1.9	Sources . . . . .	100
<b>8</b>	<b>Statistical Tests and Models</b>	<b>101</b>
8.1	Tests for Exploratory Data Analysis . . . . .	101
8.2	Statistical Modeling . . . . .	103
8.2.1	Installation of <code>statsmodels</code> . . . . .	103
8.2.2	Linear Model . . . . .	103
8.2.3	Generalized Linear Regression . . . . .	108
8.3	Validating the Results of Logistic Regression . . . . .	112
8.3.1	Confusion Matrix . . . . .	113
8.3.2	Accuracy . . . . .	114
8.3.3	Precision . . . . .	115
8.3.4	Recall . . . . .	115
8.3.5	F-beta Score . . . . .	116
8.3.6	Receiver Operating Characteristic (ROC) Curve . . . . .	117
8.3.7	Demonstration . . . . .	117
8.4	LASSO Logistic Models . . . . .	121
8.4.1	Theoretical Formulation of the Problem . . . . .	121
8.4.2	Solution Path . . . . .	122
8.4.3	Selection the Tuning Parameter . . . . .	126
<b>9</b>	<b>Supervised Learning</b>	<b>129</b>
9.1	Introduction . . . . .	129
9.2	Classification vs Regression . . . . .	129
9.2.1	Regression metrics . . . . .	129
9.2.2	Classification metrics . . . . .	130
9.2.3	Cross-validation . . . . .	131
<b>10</b>	<b>Exercises</b>	<b>133</b>
	<b>References</b>	<b>141</b>

# Preliminaries

The notes were developed with Quarto; for details about Quarto, visit <https://quarto.org/docs/books>.

This book is free and is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 United States License.

## Sources at GitHub

These lecture notes for STAT 3255/5255 in Fall 2024 represent a collaborative effort between Professor Jun Yan and the students enrolled in the course. This cooperative approach to education was facilitated through the use of GitHub, a platform that encourages collaborative coding and content development. To view these contributions and the lecture notes in their entirety, please visit our Spring 2024 repository at <https://github.com/statds/ids-f24>.

Students contributed to the lecture notes by submitting pull requests to our dedicated GitHub repository. This method not only enriched the course material but also provided students with practical experience in collaborative software development and version control.

For those interested in exploring the lecture notes from the previous years, the Spring 2024, Spring 2023 and Spring 2022 are both publicly accessible. These archives offer valuable insights into the evolution of the course content and the different perspectives brought by successive student cohorts.

*Preliminaries*

## **Midterm Project**

TBA

## **Final Project**

Students are encouraged to start designing their final projects from the beginning of the semester. There are many open data that can be used. Here is a list of data challenges that you may find useful:

- ASA Data Challenge Expo
- Kaggle
- DrivenData
- Top 10 Data Science Competitions in 2024

If you work on sports analytics, you are welcome to submit a poster to UConn Sports Analytics Symposium (UCSAS) 2024.

## **Adapting to Rapid Skill Acquisition**

In this course, students are expected to rapidly acquire new skills, a critical aspect of data science. To emphasize this, consider this insightful quote from VanderPlas (2016):

When a technologically-minded person is asked to help a friend, family member, or colleague with a computer problem, most of the time it's less a matter of knowing the answer as much as knowing how to quickly find an unknown answer. In data science it's the same: searchable web resources such as on-line documentation, mailing-list threads, and StackOverflow answers contain a wealth of information, even (especially?) if



it is a topic you've found yourself searching before. Being an effective practitioner of data science is less about memorizing the tool or command you should use for every possible situation, and more about learning to effectively find the information you don't know, whether through a web search engine or another means.

This quote captures the essence of what we aim to develop in our students: the ability to swiftly navigate and utilize the vast resources available to solve complex problems in data science. Examples tasks are: install needed software (or even hardware); search and find solutions to encountered problems.

## **Wishlist**

This is a wish list from all members of the class (alphabetical order, last name first, comma, then first name). Here is an example.

- Yan, Jun
  - Make practical data science tools accessible to undergraduates
  - Co-develop a Quarto book in collaboration with the students
  - Train students to participate real data science competitions

Add yours through a pull request; note the syntax of nested list in Markdown.

- Akach, Suha
  - Challenge and push myself to be better at python and all its libraries.
  - Be confident in my abilities of programming and making statistical inferences that are correct.
  - Be able to create my own personal project in class on time.

## *Preliminaries*

- Astle, Jaden
  - I’ve used Git before, but I’d like to become more comfortable using it and get more used to different issues that arise.
  - I’d like to learn more effective ways to “tell the story” of data analysis and show empowering visualizations.
  - I’d like to explore more methods that professional data scientists use in their model trainings to share with UConn’s Data Science Club.
- Babiec, Owen
  - Become more comfortable with Git and Github and their applications
  - Better understand the Data Science pipeline and workflow
  - Learn how to show my skills I have learned in this class during interviews
- Baptista, Stef
  - Develop a project/presentation suitable enough for industry
  - Improve on my data science skills regarding pandas and numpy
  - Understanding the scope of packages in python as a language
- Bienvenue, Jack
  - Learn professional visualization techniques, particularly for geospatial data
  - Foster a high level of working knowledge of Git
  - Create a small portfolio of examples and projects for later reference
- Blanchard, Zachary
  - Gain experience working and collaborating on projects in Git
  - Improve computer programming skills and familiarity with Python
  - Teach other students about creating presentations using Quarto

## *Wishlist*

- Borowski, Emily
  - Gain a greater understanding of Quarto and GitHub
  - Become more comfortable with my coding abilities
  - Acquire a deeper understanding of data science
- Clokey, Sara
  - Become more familiar with GitHub and Quarto
  - Execute a data science project from start to finish
- Desroches, Melanie
  - Explore the field of data science as a possible future career
  - Develop data science and machine learning skills
  - Become better at programming with Python and using Git/GitHub
- Febles, Xavier
  - Gain a further understanding of GitHub
  - Develop data visualization skills
  - Learn applications of skills learned in previous courses
- Jha, Aansh
  - Be a better student of the data science field
  - Hone git to work in colabarative workspaces
  - Learn better methods in data visualization
- Johnson, Dorothea
  - Enter data science contests
  - Familiarize myself with using Python for data Science
  - Develop a proficiency in Github
- Kashalapov, Olivia
  - Better understand neural networks
  - Machine learning utilizing Python

## *Preliminaries*

- Creating and analyzing predictive models for informed decision making
- Manna, Rahul
  - Use knowledge gained and skills developed in class to study real-world problems such as climate change.
  - Obtain a basic understanding of machine learning
- Mazzola, Julia
  - Become proficient in Git and Github.
  - Have a better understanding of data science best practices and techniques.
  - Deepen my knowledge of Python programming concepts and libraries.
- Paricharak, Aditya
  - Master Commandline Interface
  - Apply my statistical knowladge and skills to course work
  - Understand how to work with datasets
- Parvez, Mohammad Shahriyar
  - Familiarizing myself with GitHub to effectively track and manage the entire data analysis process.
  - Adopting Quarto for improved documentation of my data workflows.
  - Exploring advanced techniques for data analysis and visualization.
  - Developing my personal Git repository and publishing data projects as a professional website.
- Tan, Qianruo
  - Learn how to use GitHub, and create my own page
  - Get a good grade on this class

- Learn more about how to processing data
- Xu, Deyu

## Presentation Orders

The topic presentation order is set up in class.

```
with open('rosters/3255.txt', 'r') as file:
    ug = [line.strip() for line in file]
with open('rosters/5255.txt', 'r') as file:
    gr = [line.strip() for line in file]
presenters = ug + gr
target = "Blanchard" # pre-arranged 1st presenter
presenters = [name for name in presenters if target not in name]

import random
## seed jointly set by the class
random.seed(5347 + 2896 + 9050 + 1687 + 63)
random.sample(presenters, len(presenters))
## random.shuffle(presenters) # This would shuffle the list in place
```

```
['Xu,Deyu',
 'Clokey,Sara Karen',
 'Johnson,Dorothea Trixie',
 'Febles,Xavier Milan',
 'Cai,Yizhan',
 'Bienvenue,Jack Noel',
 'Mazzola,Julia Cecelia',
 'Akach,Suha',
 'Manna,Rahul',
 'Astile,Jaden Bryce',
```

## *Preliminaries*

```
'Kashalapov,Olivia',  
'Borowski,Emily Helen',  
'Tan,Qianruo',  
'Desroches,Melanie',  
'Paricharak,Aditya Sushant',  
'Jha,Aansh',  
'Babiec,Owen Thomas',  
'Baptista,Stef Clare',  
'Parvez,Mohammad Shahriyar']
```

Switching slots is allowed as long as you find someone who is willing to switch with you. In this case, make a pull request to switch the order and let me know.

You are welcome to choose a topic that you are interested the most, subject to some order restrictions. For example, decision tree should be presented before random forest or extreme gradient boosting. This justifies certain requests for switching slots.

## **Course Logistics**

### **Presentation Task Board**

Here are some example tasks:

- Making presentations with Quarto
- Data science ethics
- Data science communication skills
- Import/Export data
- Arrow as a cross-platform data format
- Database operation with Structured query language (SQL)
- Grammar of graphics
- Handling spatial data

## Course Logistics

- Visualize spatial data in a Google map
- Animation
- Classification and regression trees
- Support vector machine
- Random forest
- Naive Bayes
- Bagging vs boosting
- Neural networks
- Deep learning
- TensorFlow
- Autoencoders
- Reinforcement learning
- Calling C/C++ from Python
- Calling R from Python and vice versa
- Developing a Python package

Please use the following table to sign up.

Date	Presenter	Topic
09/11	Zachary Blanchard	Making presentation with Quarto
09/16	Deyu Xu	Import/Export data
09/18	Sara Clokey	Communications in data science
09/23	Doratheia Johnson	Database with SQL
09/25	Xavier Febles	

### *Preliminaries*

Date	Presenter	Topic
09/30	Jack Bienvenue	Visualizing spatial data in a Google Map
10/02	Julia Mazzola	Data Visualization with Plotnine
10/07	Suha Akach	Career readiness in data science
10/09	Rahul Manna	
10/14	Kaitlyn Anderson	
10/16	TBA	Classification and regression trees
10/21	Jaden Astle	
10/23	Olivia Kashalapov	Random Forest
10/28	Emily Borowski	
10/30	Qianruo Tan	Reinforcement Learning
11/04	Melanie Desroches	Calling R from Python and vice versa
11/06	Aditya Paricharak	
11/11	Aansh Jha	
11/11	Owen Babiec	
11/13	Stef Baptista	



Date	Presenter	Topic
11/13	Mohammad Parvez	Developing a Python package

## Final Project Presentation Schedule

We use the same order as the topic presentation for undergraduate final presentation. An introduction on how to use Quarto to prepare presentation slides is available under the `templates` directory in the classnotes source tree, thank to Zachary Blanchard, which can be used as a template to start with.

Date	Presenter
11/18	Sara Clokey; Dorothea Johnson; Xavier Febles; Jack Bienvenue
11/20	Julia Mazzola; Suha Akach; Rahul Manna; Jaden Astle
12/02	Olivia Kashalapov; Emily Borowski Qianruo Tan; Melanie Desroches
12/04	Aditya Paricharak; Aansh Jha; Owen Babiec; Stef Baptista

## Contributing to the Class Notes

Contribution to the class notes is through a ‘pull request’.

- Start a new branch and switch to the new branch.
- On the new branch, add a `qmd` file for your presentation
- If using Python, create and activate a virtual environment with `requirements.txt`

## Preliminaries

- Edit `_quarto.yml` add a line for your `qmd` file to include it in the notes.
- Work on your `qmd` file, test with `quarto render`.
- When satisfied, commit and make a pull request with your quarto files and an updated `requirements.txt`.

I have added a template file `mysection.qmd` and a new line to `_quarto.yml` as an example.

For more detailed style guidance, please see my notes on statistical writing.

Plagiarism is to be prevented. Remember that these class notes are publicly available online with your names attached. Here are some resources on [how to avoid plagiarism](<https://usingsources.fas.harvard.edu/how-avoid-plagiarism>). In particular, in our course, one convenient way to avoid plagiarism is to use our own data (e.g., NYC Open Data). Combined with your own explanation of the code chunks, it would be hard to plagiarize.

## Homework Requirements

- Use the repo from Git Classroom to submit your work. See Section Chapter 2.
  - Keep the repo clean (no tracking generated files).
    - \* Never “Upload” your files; use the git command lines.
    - \* Make commit message informative (think about the readers).
- Use `quarto` source only. See Chapter 3.
- For the convenience of grading, add your html output to a release in your repo.
- For standalone pdf output, you will need to have LaTeX installed.

## **Practical Tips**

### **Data analysis**

- Use an IDE so you can play with the data interactively
- Collect codes that have tested out into a script for batch processing
- During data cleaning, keep in mind how each variable will be used later
- No keeping large data files in a repo; assume a reasonable location with your collaborators

### **Presentation**

- Don't forget to introduce yourself if there is no moderator.
- Highlight your research questions and results, not code.
- Give an outline, carry it out, and summarize.
- Use your own examples to reduce the risk of plagiarism.

## **My Presentation Topic (Template)**

### **Introduction**

Put an overview here. Use Markdown syntax.

### **Sub Topic 1**

Put materials on topic 1 here

Python examples can be put into `python` code chunks:

## *Preliminaries*

```
import pandas as pd  
  
# do something
```

## **Sub Topic 2**

Put materials on topic 2 here.

## **Sub Topic 3**

Put materials on topic 3 here.

## **Conclusion**

Put summaries here.

# 1 Introduction

## 1.1 What Is Data Science?

Data science is a multifaceted field, often conceptualized as resting on three fundamental pillars: mathematics/statistics, computer science, and domain-specific knowledge. This framework helps to underscore the interdisciplinary nature of data science, where expertise in one area is often complemented by foundational knowledge in the others.

A compelling definition was offered by Prof. Bin Yu in her 2014 Presidential Address to the Institute of Mathematical Statistics. She defines

$$\text{Data Science} = \text{SDC}^3,$$

where

- ‘S’ represents Statistics, signifying the crucial role of statistical methods in understanding and interpreting data;
- ‘D’ stands for domain or science knowledge, indicating the importance of specialized expertise in a particular field of study;
- the three ‘C’s’ denotes computing, collaboration/teamwork, and communication to outsiders.

Computing underscores the need for proficiency in programming and algorithmic thinking, collaboration/teamwork reflects the inherently collaborative nature of data science projects, often requiring teams with diverse skill sets, and communication to outsiders emphasizes the importance of

## 1 Introduction

translating complex data insights into understandable and actionable information for non-experts.

This definition neatly captures the essence of data science, emphasizing a balance between technical skills, teamwork, and the ability to communicate effectively.

## 1.2 Expectations from This Course

In this course, students will be expected to achieve the following outcomes:

- **Proficiency in Project Management with Git:** Develop a solid understanding of Git for efficient and effective project management. This involves mastering version control, branching, and collaboration through this powerful tool.
- **Proficiency in Project Reporting with Quarto:** Gain expertise in using Quarto for professional-grade project reporting. This encompasses creating comprehensive and visually appealing reports that effectively communicate your findings.
- **Hands-On Experience with Real-World Data Science Projects:** Engage in practical data science projects that reflect real-world scenarios. This hands-on approach is designed to provide you with direct experience in tackling actual data science challenges.
- **Competency in Using Python and Its Extensions for Data Science:** Build strong skills in Python, focusing on its extensions relevant to data science. This includes libraries like Pandas, NumPy, and Matplotlib, among others, which are critical for data analysis and visualization.

## 1.3 Computing Environment

- **Full Grasp of the Meaning of Results from Data Science Algorithms:** Learn to not only apply data science algorithms but also to deeply understand the implications and meanings of their results. This is crucial for making informed decisions based on these outcomes.
- **Basic Understanding of the Principles of Data Science Methods:** Acquire a foundational knowledge of the underlying principles of various data science methods. This understanding is key to effectively applying these methods in practice.
- **Commitment to the Ethics of Data Science:** Emphasize the importance of ethical considerations in data science. This includes understanding data privacy, bias in data and algorithms, and the broader social implications of data science work.

## 1.3 Computing Environment

All setups are operating system dependent. As soon as possible, stay away from Windows. Otherwise, good luck (you will need it).

### 1.3.1 Command Line Interface

On Linux or MacOS, simply open a terminal.

On Windows, several options can be considered.

- Windows Subsystem Linux (WSL): <https://learn.microsoft.com/en-us/windows/wsl/>
- Cygwin (with X): <https://x.cygwin.com>
- Git Bash: <https://www.gitkraken.com/blog/what-is-git-bash>

## 1 Introduction

To jump start, here is a tutorial: Ubuntu Linux for beginners.

At least, you need to know how to handle files and traverse across directories. The tab completion and introspection supports are very useful.

Here are several commonly used shell commands:

- **cd**: change directory; `..` means parent directory.
- **pwd**: present working directory.
- **ls**: list the content of a folder; `-l` long version; `-a` show hidden files; `-t` ordered by modification time.
- **mkdir**: create a new directory.
- **cp**: copy file/folder from a source to a target.
- **mv**: move file/folder from a source to a target.
- **rm**: remove a file a folder.

### 1.3.2 Python

Set up Python on your computer:

- Python 3.
- Python package manager **miniconda** or **pip**.
- Integrated Development Environment (IDE) (Jupyter Notebook; RStudio; VS Code; Emacs; etc.)

I will be using VS Code in class.

Readability is important! Check your Python coding styles against the recommended styles: <https://peps.python.org/pep-0008/>. A good place to start is the Section on “Code Lay-out”.

Online books on Python for data science:

- “Python Data Science Handbook: Essential Tools for Working with Data,” First Edition, by Jake VanderPlas, O’Reilly Media, 2016.



### *1.3 Computing Environment*

2. “Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.” Third Edition, by Wes McKinney, O’Reilly Media, 2022.



## 2 Project Management

Many tutorials are available in different formats. Here is a YouTube video “Git and GitHub for Beginners — Crash Course’’. The video also covers GitHub, a cloud service for Git which provides a cloud back up of your work and makes collaboration with co-workers easy. Similar services are, for example, bitbucket and GitLab.

There are tools that make learning Git easy.

- Here is a collection of online Git exercises that I used for Git training in other courses that I taught.
- Here is a game called *Oh My Git*, an open source game about learning Git!

### 2.1 Set Up Git/GitHub

Download Git if you don’t have it already.

To set up GitHub (other services like Bitbucket or GitLab are similar), you need to

- Generate an SSH key if you don’t have one already.
- Sign up an GitHub account.
- Add the SSH key to your GitHub account

See how to get started with GitHub account.

## 2.2 Most Frequently Used Git Commands

- `git clone`:
  - Used to clone a repository to a local folder.
  - Requires either HTTPS link or SSH key to authenticate.
- `git pull`:
  - Downloads any updates made to the remote repository and automatically updates the local repository.
- `git status`:
  - Returns the state of the working directory.
  - Lists the files that have been modified, and are yet to be or have been staged and/or committed.
  - Shows if the local repository is behind or ahead a remote branch.
- `git add`:
  - Adds new or modified files to the Git staging area.
  - Gives the option to select which files are to be sent to the remote repository
- `git rm`:
  - Used to remove files from the staging index or the local repository.
- `git commit`:
  - Commits changes made to the local repository and saves it like a snapshot.
  - A message is recommended with every commit to keep track of changes made.
- `git push`:
  - Used to send commits made on local repository to the remote repository.

## 2.3 Tips on using Git:

- Use the command line interface instead of the web interface (e.g., upload on GitHub)
- Make frequent small commits instead of rare large commits.
- Make commit messages informative and meaningful.
- Name your files/folders by some reasonable convention.
  - Lower cases are better than upper cases.
  - No blanks in file/folder names.
- Keep the repo clean by not tracking generated files.
- Create a `.gitignore` file for better output from `git status`.
- Keep the linewidth of sources to under 80 for better `git diff` view.

## 2.4 Pull Request

To contribute to an open source project (e.g., our classnotes), use pull requests. Pull requests “let you tell others about changes you’ve pushed to a branch in a repository on GitHub. Once a pull request is opened, you can discuss and review the potential changes with collaborators and add follow-up commits before your changes are merged into the base branch.”

Watch this YouTube video: GitHub pull requests in 100 seconds.



## 3 Reproducible Data Science

Data science projects should be reproducible to be trustworthy. Dynamic documents facilitate reproducibility. Quarto is an open-source dynamic document preparation system, ideal for scientific and technical publishing. From the official websites, Quarto can be used to:

- Create dynamic content with Python, R, Julia, and Observable.
- Author documents as plain text markdown or Jupyter notebooks.
- Publish high-quality articles, reports, presentations, websites, blogs, and books in HTML, PDF, MS Word, ePub, and more.
- Author with scientific markdown, including equations, citations, cross references, figure panels, callouts, advanced layout, and more.

### 3.1 Introduction to Quarto

To get started with Quarto, see documentation at [Quarto](#).

For a clean style, I suggest that you use `VS Code` as your IDE. The `ipynb` files have extra formats in plain texts, which are not as clean as `qmd` files. There are, of course, tools to convert between the two representations of a notebook. For example:

```
quarto convert hello.ipynb # converts to qmd
quarto convert hello.qmd   # converts to ipynb
```

### 3 Reproducible Data Science

We will use Quarto for homework assignments, classnotes, and presentations. You will see them in action through in-class demonstrations. The following sections in the Quarto Guide are immediately useful.

- Markdown basics
- Using Python
- Presentations

A template for homework is in this repo (`hwtemp.qmd`) to get you started with homework assignments.

## 3.2 Compiling the Classnotes

The sources of the classnotes are at <https://github.com/statds/ids-f24>. This is also the source tree that you will contribute to this semester. I expect that you clone the repository to your own computer, update it frequently, and compile the latest version on your computer (reproducibility).

To compile the classnotes, you need the following tools: Git, Quarto, and Python.

### 3.2.1 Set up your Python Virtual Environment

I suggest that a Python virtual environment for the classnotes be set up in the current directory for reproducibility. A Python virtual environment is simply a directory with a particular file structure, which contains a specific Python interpreter and software libraries and binaries needed to support a project. It allows us to isolate our Python development projects from our system installed Python and other Python environments.

To create a Python virtual environment for our classnotes:



## 3.2 Compiling the Classnotes

```
python3 -m venv .ids-f24-venv
```

Here `.ids-f24-venv` is the name of the virtual environment to be created. Choose an informative name. This only needs to be set up once.

To activate this virtual environment:

```
.. .ids-f24-venv/bin/activate
```

After activating the virtual environment, you will see `(.ids-f24-venv)` at the beginning of your shell prompt. Then, the Python interpreter and packages needed will be the local versions in this virtual environment without interfering your system-wide installation or other virtual environments.

To install the Python packages that are needed to compile the classnotes, we have a `requirements.txt` file that specifies the packages and their versions. They can be installed easily with:

```
pip install -r requirements.txt
```

If you are interested in learning how to create the `requirements.txt` file, just put your question into a Google search.

To exit the virtual environment, simply type `deactivate` in your command line. This will return you to your system's global Python environment.

### 3.2.2 Clone the Repository

Clone the repository to your own computer. In a terminal (command line), go to an appropriate directory (folder), and clone the repo. For example, if you use `ssh` for authentication:

```
git clone git@github.com:statds/ids-f24.git
```

### *3 Reproducible Data Science*

#### **3.2.3 Render the Classnotes**

Assuming `quarto` has been set up, we render the classnotes in the cloned repository

```
cd ids-f24
quarto render
```

If there are error messages, search and find solutions to clear them. Otherwise, the html version of the notes will be available under `_book/index.html`, which is default location of the output.

## 4 Python Refreshment

### 4.1 Know Your Computer

#### 4.1.1 Operating System

Your computer has an operating system (OS), which is responsible for managing the software packages on your computer. Each operating system has its own package management system. For example:

- **Linux:** Linux distributions have a variety of package managers depending on the distribution. For instance, Ubuntu uses APT (Advanced Package Tool), Fedora uses DNF (Dandified Yum), and Arch Linux uses Pacman. These package managers are integral to the Linux experience, allowing users to install, update, and manage software packages easily from repositories.
- **macOS:** macOS uses Homebrew as its primary package manager. Homebrew simplifies the installation of software and tools that aren't included in the standard macOS installation, using simple commands in the terminal.
- **Windows:** Windows users often rely on the Microsoft Store for apps and software. For more developer-focused package management, tools like Chocolatey and Windows Package Manager (Winget) are used. Additionally, recent versions of Windows have introduced the Windows Subsystem for Linux (WSL). WSL allows Windows users to run a Linux environment directly on Windows, unifying Windows

## 4 Python Refreshment

and Linux applications and tools. This is particularly useful for developers and data scientists who need to run Linux-specific software or scripts. It saves a lot of trouble Windows users used to have before its time.

Understanding the package management system of your operating system is crucial for effectively managing and installing software, especially for data science tools and applications.

### 4.1.2 File System

A file system is a fundamental aspect of a computer's operating system, responsible for managing how data is stored and retrieved on a storage device, such as a hard drive, SSD, or USB flash drive. Essentially, it provides a way for the OS and users to organize and keep track of files. Different operating systems typically use different file systems. For instance, NTFS and FAT32 are common in Windows, APFS and HFS+ in macOS, and Ext4 in many Linux distributions. Each file system has its own set of rules for controlling the allocation of space on the drive and the naming, storage, and access of files, which impacts performance, security, and compatibility. Understanding file systems is crucial for tasks such as data recovery, disk partitioning, and managing file permissions, making it an important concept for anyone working with computers, especially in data science and IT fields.

Navigating through folders in the command line, especially in Unix-like environments such as Linux or macOS, and Windows Subsystem for Linux (WSL), is an essential skill for effective file management. The command `cd` (change directory) is central to this process. To move into a specific directory, you use `cd` followed by the directory name, like `cd Documents`. To go up one level in the directory hierarchy, you use `cd ..`. To return to the home directory, simply typing `cd` or `cd ~` will suffice. The `ls` command lists all files and folders in the current directory, providing a clear view of your options for navigation. Mastering these commands,

## 4.2 The Python World

along with others like `pwd` (print working directory), which displays your current directory, equips you with the basics of moving around the file system in the command line, an indispensable skill for a wide range of computing tasks in Unix-like systems.

You have programmed in Python. Regardless of your skill level, let us do some refreshing.

## 4.2 The Python World

- **Function:** a block of organized, reusable code to complete certain task.
- **Module:** a file containing a collection of functions, variables, and statements.
- **Package:** a structured directory containing collections of modules and an `__init.py__` file by which the directory is interpreted as a package.
- **Library:** a collection of related functionality of codes. It is a reusable chunk of code that we can use by importing it in our program, we can just use it by importing that library and calling the method of that library with `period(.)`.

See, for example, how to build a Python library.

## 4.3 Standard Library

Python's has an extensive standard library that offers a wide range of facilities as indicated by the long table of contents listed below. See documentation online.

## 4 Python Refreshment

The library contains built-in modules (written in C) that provide access to system functionality such as file I/O that would otherwise be inaccessible to Python programmers, as well as modules written in Python that provide standardized solutions for many problems that occur in everyday programming. Some of these modules are explicitly designed to encourage and enhance the portability of Python programs by abstracting away platform-specifics into platform-neutral APIs.

Question: How to get the constant  $e$  to an arbitrary precision?

The constant is only represented by a given double precision.

```
import math
print("%0.20f" % math.e)
print("%0.80f" % math.e)
```

2.71828182845904509080

2.71828182845904509079559829842764884233474731445312500000000000000000000000

Now use package `decimal` to export with an arbitrary precision.

```
import decimal # for what?

## set the required number digits to 150
decimal.getcontext().prec = 150
decimal.Decimal(1).exp().to_eng_string()
decimal.Decimal(1).exp().to_eng_string()[2:]
```

'718281828459045235360287471352662497757247093699959574966967627724076630353

## 4.4 Important Libraries

- NumPy
- pandas
- matplotlib
- IPython/Jupyter
- SciPy
- scikit-learn
- statsmodels

Question: how to draw a random sample from a normal distribution and evaluate the density and distributions at these points?

```
from scipy.stats import norm

mu, sigma = 2, 4
mean, var, skew, kurt = norm.stats(mu, sigma, moments='mvsk')
print(mean, var, skew, kurt)
x = norm.rvs(loc = mu, scale = sigma, size = 10)
x
```

```
2.0 16.0 0.0 0.0
```

```
array([ 5.06579271,  4.54645018,  1.06566456,  7.96340825,  2.98506217,
        -2.37312315, -0.89449371,  0.64125325,  2.00906353,  0.46073374])
```

The pdf and cdf can be evaluated:

```
norm.pdf(x, loc = mu, scale = sigma)
```

```
array([0.07435129, 0.08144147, 0.09705149, 0.03282539, 0.09675665,
        0.05486584, 0.07676179, 0.09414431, 0.09973531, 0.09261772])
```

## 4.5 Writing a Function

Consider the Fibonacci Sequence 1, 1, 2, 3, 5, 8, 13, 21, 34, .... The next number is found by adding up the two numbers before it. We are going to use 3 ways to solve the problems.

The first is a recursive solution.

```
def fib_rs(n):
    if (n==1 or n==2):
        return 1
    else:
        return fib_rs(n - 1) + fib_rs(n - 2)

%timeit fib_rs(10)
```

9.05 s ± 118 ns per loop (mean ± std. dev. of 7 runs, 100,000 loops each)

The second uses dynamic programming memoization.

```
def fib_dm_helper(n, mem):
    if mem[n] is not None:
        return mem[n]
    elif (n == 1 or n == 2):
        result = 1
    else:
        result = fib_dm_helper(n - 1, mem) + fib_dm_helper(n - 2, mem)
    mem[n] = result
    return result

def fib_dm(n):
    mem = [None] * (n + 1)
    return fib_dm_helper(n, mem)
```



## 4.5 Writing a Function

```
%timeit fib_dm(10)
```

1.98 s  $\pm$  29 ns per loop (mean  $\pm$  std. dev. of 7 runs, 1,000,000 loops each)

The third is still dynamic programming but bottom-up.

```
def fib_dbu(n):
    mem = [None] * (n + 1)
    mem[1] = 1;
    mem[2] = 1;
    for i in range(3, n + 1):
        mem[i] = mem[i - 1] + mem[i - 2]
    return mem[n]
```

```
%timeit fib_dbu(500)
```

74.1 s  $\pm$  7.21 s per loop (mean  $\pm$  std. dev. of 7 runs, 10,000 loops each)

Apparently, the three solutions have very different performance for larger  $n$ .

### 4.5.1 Monte Hall

Here is a function that performs the Monte Hall experiments.

```
import numpy as np

def montehall(ndoors, ntrials):
    doors = np.arange(1, ndoors + 1) / 10
```

## 4 Python Refreshment

```
prize = np.random.choice(doors, size=ntrials)
player = np.random.choice(doors, size=ntrials)
host = np.array([np.random.choice([d for d in doors
                                   if d not in [player[x], prize[x]]])
                 for x in range(ntrials)])
player2 = np.array([np.random.choice([d for d in doors
                                      if d not in [player[x], host[x]]])
                   for x in range(ntrials)])
return {'noswitch': np.sum(prize == player), 'switch': np.sum(prize == p
```

Test it out:

```
montehall(3, 1000)
montehall(4, 1000)
```

```
{'noswitch': np.int64(242), 'switch': np.int64(365)}
```

The true value for the two strategies with  $n$  doors are, respectively,  $1/n$  and  $\frac{n-1}{n(n-2)}$ .

## 4.6 Variables versus Objects

In Python, variables and the objects they point to actually live in two different places in the computer memory. Think of variables as pointers to the objects they're associated with, rather than being those objects. This matters when multiple variables point to the same object.

```
x = [1, 2, 3] # create a list; x points to the list
y = x        # y also points to the same list in the memory
y.append(4)  # append to y
x           # x changed!
```

## 4.6 Variables versus Objects

```
[1, 2, 3, 4]
```

Now check their addresses

```
print(id(x))    # address of x
print(id(y))    # address of y
```

```
4758266496
4758266496
```

Nonetheless, some data types in Python are “immutable”, meaning that their values cannot be changed in place. One such example is strings.

```
x = "abc"
y = x
y = "xyz"
x
```

```
'abc'
```

Now check their addresses

```
print(id(x))    # address of x
print(id(y))    # address of y
```

```
4507939064
4574102032
```

## 4 Python Refreshment

Question: What's mutable and what's immutable?

Anything that is a collection of other objects is mutable, except **tuples**.

Not all manipulations of mutable objects change the object rather than create a new object. Sometimes when you do something to a mutable object, you get back a new object. Manipulations that change an existing object, rather than create a new one, are referred to as “in-place mutations” or just “mutations.” So:

- **All** manipulations of immutable types create new objects.
- **Some** manipulations of mutable types create new objects.

Different variables may all be pointing at the same object is preserved through function calls (a behavior known as “pass by object-reference”). So if you pass a list to a function, and that function manipulates that list using an in-place mutation, that change will affect any variable that was pointing to that same object outside the function.

```
x = [1, 2, 3]
y = x

def append_42(input_list):
    input_list.append(42)
    return input_list

append_42(x)
```

[1, 2, 3, 42]

Note that both `x` and `y` have been appended by 42.

## 4.7 Number Representation

Numbers in a computer's memory are represented by binary styles (on and off of bits).

### 4.7.1 Integers

If not careful, It is easy to be bitten by overflow with integers when using Numpy and Pandas in Python.

```
import numpy as np

x = np.array(2 ** 63 - 1 , dtype = 'int')
x
# This should be the largest number numpy can display, with
# the default int8 type (64 bits)
```

```
array(9223372036854775807)
```

*Note: on Windows and other platforms, `dtype = 'int'` may have to be changed to `dtype = np.int64` for the code to execute. Source: Stackoverflow*

What if we increment it by 1?

```
y = np.array(x + 1, dtype = 'int')
y
# Because of the overflow, it becomes negative!
```

```
array(-9223372036854775808)
```

## 4 Python Refreshment

For vanilla Python, the overflow errors are checked and more digits are allocated when needed, at the cost of being slow.

```
2 ** 63 * 1000
```

```
9223372036854775808000
```

This number is 1000 times larger than the prior number, but still displayed perfectly without any overflows

### 4.7.2 Floating Number

Standard double-precision floating point number uses 64 bits. Among them, 1 is for sign, 11 is for exponent, and 52 are fraction significand, See [https://en.wikipedia.org/wiki/Double-precision\\_floating-point\\_format](https://en.wikipedia.org/wiki/Double-precision_floating-point_format). The bottom line is that, of course, not every real number is exactly representable.

If you have played the Game 24, here is a tricky one:

```
8 / (3 - 8 / 3) == 24
```

False

Surprise?

There are more.

```
0.1 + 0.1 + 0.1 == 0.3
```

False

## 4.7 Number Representation

```
0.3 - 0.2 == 0.1
```

False

What is really going on?

```
import decimal
decimal.Decimal(0.1)
```

```
Decimal('0.1000000000000000055511151231257827021181583404541015625')
```

```
decimal.Decimal(8 / (3 - 8 / 3))
```

```
Decimal('23.99999999999999989341858963598497211933135986328125')
```

Because the mantissa bits are limited, it can not represent a floating point that's both very big and very precise. Most computers can represent all integers up to  $2^{53}$ , after that it starts skipping numbers.

```
2.1 ** 53 + 1 == 2.1 ** 53
```

```
# Find a number larger than 2 to the 53rd
```

True

```
x = 2.1 ** 53
for i in range(1000000):
    x = x + 1
x == 2.1 ** 53
```

## 4 Python Refreshment

True

We add 1 to `x` by 1000000 times, but it still equal to its initial value, 2.1  
**\*\* 53.** This is because this number is too big that computer can't handle it with precision like add 1.

Machine epsilon is the smallest positive floating-point number `x` such that `1 + x != 1`.

```
print(np.finfo(float).eps)
print(np.finfo(np.float32).eps)
```

2.220446049250313e-16

1.1920929e-07

## 4.8 Virtual Environment

Virtual environments in Python are essential tools for managing dependencies and ensuring consistency across projects. They allow you to create isolated environments for each project, with its own set of installed packages, separate from the global Python installation. This isolation prevents conflicts between project dependencies and versions, making your projects more reliable and easier to manage. It's particularly useful when working on multiple projects with differing requirements, or when collaborating with others who may have different setups.

To set up a virtual environment, you first need to ensure that Python is installed on your system. Most modern Python installations come with the `venv` module, which is used to create virtual environments. Here's how to set one up:

- Open your command line interface.
- Navigate to your project directory.



## 4.8 Virtual Environment

- Run `python3 -m venv myenv`, where `myenv` is the name of the virtual environment to be created. Choose an informative name.

This command creates a new directory named `myenv` (or your chosen name) in your project directory, containing the virtual environment.

To start using this environment, you need to activate it. The activation command varies depending on your operating system:

- On Windows, run `myenv\Scripts\activate`.
- On Linux or MacOS, use `source myenv/bin/activate` or `. myenv/bin/activate`.

Once activated, your command line will typically show the name of the virtual environment, and you can then install and use packages within this isolated environment without affecting your global Python setup.

To exit the virtual environment, simply type `deactivate` in your command line. This will return you to your system's global Python environment.

As an example, let's install a package, like `numpy`, in this newly created virtual environment:

- Ensure your virtual environment is activated.
- Run `pip install numpy`.

This command installs the `requests` library in your virtual environment. You can verify the installation by running `pip list`, which should show `requests` along with its version.



# 5 Communication in Data Science

This chapter was written by Sara Clokey.

## 5.1 Introduction

Hi! My name is Sara, and I am a junior double majoring in Applied Data Analysis and Communication. The topic of my presentation today, Communication in Data Science, combines my academic and professional interests while underscoring the importance of ‘soft skills’ like public speaking, for example, within STEM fields like data science.

## 5.2 Importance & Application of Communication

Data science as a career path has exploded within the last decade. Some fields that offer data science positions include:

- Finance
- Healthcare
- Media production
- Sports
- Banking
- Insurance
- E-Commerce
- Energy

## 5 *Communication in Data Science*

- Manufacturing
- Transportation
- Construction

Because data science is applicable in so many industries, it is essential that data scientists have the skills and experience to communicate their work with others who do not have the same technical education. As analyzed by Radovilsky et al. (2018), job listings within the field of data science often include qualifications like “strong interpersonal skills” and “demonstrated presentation and communication ability,” highlighting the pervasive need for this skill set.

Within these industries, collaboration and teamwork are often at the forefront. Inexperience with data should not prevent your colleagues from being able to contribute to shared projects, and strong communication skills can mitigate this challenge!

### 5.3 General Communication Skills

#### 5.3.1 Verbal Communication Skills

Verbal Communication: “The use of sounds and language to relay a message” (Yer (2018))

Verbal Communication Tips:

- Make a good first impression
- Use appropriate language (jargon, metaphors)
- Prioritize brevity
- Practice beforehand
- Allow room for questions

### 5.3.2 Non-Verbal Communication Skills

Non-Verbal Communication: “Information, emotion, a movement that is expressed without words and without the help of language” (Grillo and Enesi (2022))

Non-Verbal Communication Tips:

- Utilize vocal variety (pitch, rate, volume)
- Avoid distracting hand and body movements
- Make eye contact
- Pay attention to proxemics

### 5.3.3 Visual Communication Skills

Visual Communication: “Any communication that employs one’s sense of sight to deliver a message without the usage of any verbal cues” (Fayaz (2022))

Visual Communication Tips:

- Prioritize clarity
- Use proper labeling and scaling
- Create visual contrast (colors, shapes, fonts)
- Choose the most appropriate visual representation

### 5.3.4 Written Communication Skills

Written Communication: “Any form of communication which is written and documented from the sender to the receiver” (Prabavathi and Nagasubramani (2018))

Written Communication Tips:

- Clearly state your goal with a thesis statement

- Maintain professionalism (contractions, slang)
- Proofread and utilize peer editing
- Follow a specific structure
- Balance concision with analysis

## 5.4 Communication in Data Science

Often, data scientists must communicate “technical conclusions to non-technical members” (Pragmatic (2024)); this may be colleagues in other departments, like marketing, or supervisors at the managerial level. Here are some tips for effectively communicating project results specifically in the field of data science.

### 5.4.1 Identify your Audience

Who are you sharing information with? Is it a room of data scientists like this one? Is it full of students who want to learn about data science? Is it a group of executives looking to make a funding decision?

- Consider the context and prior knowledge (technical jargon)
- Consider the motivation for listening

### 5.4.2 Utilize Data Visualization

One of the most effective methods of communicating results in data science, especially to those without technical coding knowledge, is data visualization techniques (Vandemeulebroecke et al. (2019)). Python uses the package ‘matplotlib’ to produce these visualizations, including:

- line plots
- bar plots

## 5.4 Communication in Data Science

- box plots
- histograms
- heat maps
- pie charts

These visualizations allow complex statistical projects to be simplified into a single graphic, focusing on project results and implications rather than methodology. Ensure that data visualization techniques are free of technical jargon and clearly label all visual aspects.

### 5.4.3 Focus on Data Communication Skills

The following skill sets highlight technical data communication that will be more common in projects with other data scientists to communicate about your data.

- Coding communication: Python, R, Julia, JavaScript, SQL, etc.
- Analysis communication: creating a storyline, descriptive versus diagnostic versus predictive analytics, problem identification
- Data management: collection, cleaning/transformation, storage
- Data visualization

### 5.4.4 Give Space for Questions and Feedback

Within professional spaces, data scientists should provide time for their clients, supervisors, and colleagues to ask questions about their work and subsequent findings.

- Pause for questions throughout the presentation
- Offer contact information for continued collaboration
- Provide a structure for anonymous feedback
- Schedule follow-ups if necessary

## 5 *Communication in Data Science*

Teamwork is often at the heart of data science projects within industries, and open communication makes this teamwork run much more smoothly.

### 5.5 Further Learning

While pursuing degrees in the data science field, consider taking Communication courses at UConn that can bolster your understanding and skill set. Some applicable communication courses include:

- COMM 2100: Professional Communication
- COMM 3110: Organizational Communication
- COMM 3430: Science Communication
- COMM 5655: Human-Computer Interaction
- COMM 5900: Professional Communication

Effective communication also requires practice. Here are some ways to practice these skills while earning your degree:

- Fully participate in group projects
- Seek presentation opportunities (class, conferences, etc.)
- Explain data science coursework to peers outside of your program
- Explore internship opportunities that involve collaboration with other departments



# 6 Data Manipulation

## 6.1 Introduction

Data manipulation is crucial for transforming raw data into a more analyzable format, essential for uncovering patterns and ensuring accurate analysis. This chapter introduces the core techniques for data manipulation in Python, utilizing the Pandas library, a cornerstone for data handling within Python's data science toolkit.

Python's ecosystem is rich with libraries that facilitate not just data manipulation but comprehensive data analysis. Pandas, in particular, provides extensive functionality for data manipulation tasks including reading, cleaning, transforming, and summarizing data. Using real-world datasets, we will explore how to leverage Python for practical data manipulation tasks.

By the end of this chapter, you will learn to:

- Import/export data from/to diverse sources.
- Clean and preprocess data efficiently.
- Transform and aggregate data to derive insights.
- Merge and concatenate datasets from various origins.
- Analyze real-world datasets using these techniques.

## **6.2 Import/Export Data**

This section was written by Deyu Xu, a MS student in Statistics at the time.

### **6.2.1 Subsection 1**

### **6.2.2 Subsection 2**

### **6.2.3 Extended Reading**

Give some references here.

## **6.3 SQL**

This section was written by Thea Johnson, a senior in Statistics at the time.

### **6.3.1 Subsection 1**

### **6.3.2 Subsection 2**

### **6.3.3 Extended Reading**

Give some references here.

## 6.4 NYC Crash Data

Consider a subset of the NYC Crash Data, which contains all NYC motor vehicle collisions data with documentation from NYC Open Data. We downloaded the crash data for the week of June 30, 2024, on September 16, 2024, in CSC format.

```
import pandas as pd

# Load the dataset
file_path = 'data/nyccrashes_2024w0630_by20240916.csv'
df = pd.read_csv(file_path)

# Replace column names: convert to lowercase and replace spaces with underscores
df.columns = df.columns.str.lower().str.replace(' ', '_')

# Display the first few rows of the dataset to understand its structure
df.head()
```

	crash_date	crash_time	borough	zip_code	latitude	longitude	location	on_s
0	06/30/2024	17:30	NaN	NaN	0.00000	0.00000	(0.0, 0.0)	NaN
1	06/30/2024	0:32	NaN	NaN	NaN	NaN	NaN	BEL
2	06/30/2024	7:05	BROOKLYN	11235.0	40.58106	-73.96744	(40.58106, -73.96744)	NaN
3	06/30/2024	20:47	NaN	NaN	40.76363	-73.95330	(40.76363, -73.9533)	FDF
4	06/30/2024	10:14	BROOKLYN	11222.0	40.73046	-73.95149	(40.73046, -73.95149)	GRE

Now we can do some cleaning after a quick browse.

```
# Replace invalid coordinates (latitude=0, longitude=0 or NaN) with NaN
df.loc[(df['latitude'] == 0) & (df['longitude'] == 0),
        ['latitude', 'longitude']] = pd.NA
```

## 6 Data Manipulation

```
df['latitude'] = df['latitude'].replace(0, pd.NA)
df['longitude'] = df['longitude'].replace(0, pd.NA)

# Longitude/latitude don't need double precision
df['latitude'] = df['latitude'].astype('float32', errors='ignore')
df['longitude'] = df['longitude'].astype('float32', errors='ignore')

# Drop the redundant 'location' column
df = df.drop(columns=['location'])

# Converting 'crash_date' and 'crash_time' columns into a single datetime column
df['crash_datetime'] = pd.to_datetime(df['crash_date'] + ' '
                                     + df['crash_time'], format='%m/%d/%Y %H:%M', errors='coerce')

# Drop the original 'crash_date' and 'crash_time' columns
df = df.drop(columns=['crash_date', 'crash_time'])
```

Are missing in zip code and borough always co-occur?

```
# Check if missing values in 'zip_code' and 'borough' always co-occur
# Count rows where both are missing
missing_cooccur = df[['zip_code', 'borough']].isnull().all(axis=1).sum()
# Count total missing in 'zip_code' and 'borough', respectively
total_missing_zip_code = df['zip_code'].isnull().sum()
total_missing_borough = df['borough'].isnull().sum()

# If missing in both columns always co-occur, the number of missing
# co-occurrences should be equal to the total missing in either column
missing_cooccur, total_missing_zip_code, total_missing_borough
```

```
(np.int64(541), np.int64(541), np.int64(541))
```

## 6.4 NYC Crash Data

Are there cases where `zip_code` and `borough` are missing but the geo codes are not missing? If so, fill in `zip_code` and `borough` using the geo codes by reverse geocoding.

First make sure `geopy` is installed.

```
pip install geopy
```

Now we use model `Nominatim` in package `geopy` to reverse geocode.

```
from geopy.geocoders import Nominatim
import time

# Initialize the geocoder; the `user_agent` is your identifier
# when using the service. Be mindful not to crash the server
# by unlimited number of queries, especially invalid code.
geolocator = Nominatim(user_agent="jyGeopyTry")
```

We write a function to do the reverse geocoding given latitude and longitude.

```
# Function to fill missing zip_code
def get_zip_code(latitude, longitude):
    try:
        location = geolocator.reverse((latitude, longitude), timeout=10)
        if location:
            address = location.raw['address']
            zip_code = address.get('postcode', None)
            return zip_code
        else:
            return None
    except Exception as e:
        print(f"Error: {e} for coordinates {latitude}, {longitude}")
```

## 6 Data Manipulation

```
        return None
    finally:
        time.sleep(1) # Delay to avoid overwhelming the service
```

Let's try it out:

```
# Example usage
latitude = 40.730610
longitude = -73.935242
zip_code = get_zip_code(latitude, longitude)
```

The function `get_zip_code` can then be applied to rows where zip code is missing but geocodes are not to fill the missing zip code.

Once zip code is known, figuring out `borough` is simple because valid zip codes from each borough are known.

### 6.5 Cross-platform Data Format Arrow

The CSV format (and related formats like TSV - tab-separated values) for data tables is ubiquitous, convenient, and can be read or written by many different data analysis environments, including spreadsheets. An advantage of the textual representation of the data in a CSV file is that the entire data table, or portions of it, can be previewed in a text editor. However, the textual representation can be ambiguous and inconsistent. The format of a particular column: Boolean, integer, floating-point, text, factor, etc. must be inferred from text representation, often at the expense of reading the entire file before these inferences can be made. Experienced data scientists are aware that a substantial part of an analysis or report generation is often the “data cleaning” involved in preparing the data for analysis. This can be an open-ended task — it required numerous trial-and-error iterations to create the list of different missing data representations we

## 6.5 Cross-platform Data Format **Arrow**

use for the sample CSV file and even now we are not sure we have them all.

To read and export data efficiently, leveraging the Apache **Arrow** library can significantly improve performance and storage efficiency, especially with large datasets. The IPC (Inter-Process Communication) file format in the context of Apache Arrow is a key component for efficiently sharing data between different processes, potentially written in different programming languages. Arrow's IPC mechanism is designed around two main file formats:

- **Stream Format:** For sending an arbitrary length sequence of Arrow record batches (tables). The stream format is useful for real-time data exchange where the size of the data is not known upfront and can grow indefinitely.
- **File (or Feather) Format:** Optimized for storage and memory-mapped access, allowing for fast random access to different sections of the data. This format is ideal for scenarios where the entire dataset is available upfront and can be stored in a file system for repeated reads and writes.

Apache Arrow provides a columnar memory format for flat and hierarchical data, optimized for efficient data analytics. It can be used in Python through the **pyarrow** package. Here's how you can use Arrow to read, manipulate, and export data, including a demonstration of storage savings.

First, ensure you have **pyarrow** installed on your computer (and preferably, in your current virtual environment):

```
pip install pyarrow
```

Feather is a fast, lightweight, and easy-to-use binary file format for storing data frames, optimized for speed and efficiency, particularly for IPC and data sharing between Python and R or Julia.

## 6 Data Manipulation

```
df.to_feather('data/nyccrashes_cleaned.feather')

# Compare the file sizes of the feather format and the CSV format
import os

# File paths
csv_file = 'data/nyccrashes_2024w0630_by20240916.csv'
feather_file = 'data/nyccrashes_cleaned.feather'

# Get file sizes in bytes
csv_size = os.path.getsize(csv_file)
feather_size = os.path.getsize(feather_file)

# Convert bytes to a more readable format (e.g., MB)
csv_size_mb = csv_size / (1024 * 1024)
feather_size_mb = feather_size / (1024 * 1024)

# Print the file sizes
print(f"CSV file size: {csv_size_mb:.2f} MB")
print(f"Feather file size: {feather_size_mb:.2f} MB")
```

Read the feather file back in:

```
dff = pd.read_feather("data/nyccrashes_cleaned.feather")
dff.shape
```

## 6.6 Using the Census Data

The US Census provides a lot of useful data that could be merged with the NYC crash data for further analytics.



## 6.6 Using the Census Data

First, ensure the DataFrame (df) is ready for merging with census data. Specifically, check that the `zip_code` column is clean and consistent and consistent.

```
import pandas as pd
df = pd.read_feather("data/nyccrashes_cleaned.feather")

valid_zip_df = df.dropna(subset=['zip_code']).copy()
valid_zip_df['zip_code'] = valid_zip_df['zip_code'].astype(int).astype(str).str.zfill(5)
unique_zips = valid_zip_df['zip_code'].unique()
```

We can use the `uszipcode` package to get basic demographic data for each zip code. For more detailed or specific census data, using the `CensusData` package or direct API calls to the Census Bureau's API.

The `uszipcode` package provides a range of information about ZIP codes in the United States. When you query a ZIP code using `uszipcode`, you can access various attributes related to demographic data, housing, geographic location, and more. Here are some of the key variables available at the ZIP code level:

### Demographic Information

- `population`: The total population.
- `population_density`: The population per square kilometer.
- `housing_units`: The total number of housing units.
- `occupied_housing_units`: The number of occupied housing units.
- `median_home_value`: The median value of homes.
- `median_household_income`: The median household income.
- `age_distribution`: A breakdown of the population by age.

### Geographic Information

- `zipcode`: The ZIP code.
- `zipcode_type`: The type of ZIP code (e.g., Standard, PO Box).

## 6 Data Manipulation

- `major_city`: The major city associated with the ZIP code.
- `post_office_city`: The city name recognized by the U.S. Postal Service.
- `common_city_list`: A list of common city names for the ZIP code.
- `county`: The county in which the ZIP code is located.
- `state`: The state in which the ZIP code is located.
- `lat`: The latitude of the approximate center of the ZIP code.
- `lng`: The longitude of the approximate center of the ZIP code.
- `timezone`: The timezone of the ZIP code.

### Economic and Housing Data

- `land_area_in_sqmi`: The land area in square miles.
- `water_area_in_sqmi`: The water area in square miles.
- `occupancy_rate`: The rate of occupancy for housing units.
- `median_age`: The median age of the population.

Install the `uszipcode` package into the current virtual environment, if it has not been installed yet.

```
pip install uszipcode
```

```
..
```

```
We will first clean the `zip_code` column to ensure it only
contains valid ZIP codes. Then, we will use a vectorized
approach to fetch the required data for each unique zip code
and merge this information back into the original `DataFrame`.
`{python}
```

```
# Remove rows where 'zip_code' is missing or not a valid ZIP code format
valid_zip_df = df.dropna(subset=['zip_code']).copy()
valid_zip_df['zip_code'] = valid_zip_df['zip_code'].astype(str).str.zfill(5)
```

Since `uszipcode` doesn't currently support vectorized operations for multiple ZIP code queries, we'll optimize the process by querying

## 6.6 Using the Census Data

each unique ZIP code once, then merging the results with the original `DataFrame`. This approach minimizes redundant queries for ZIP codes that appear multiple

times.

```
from uszipcode import SearchEngine

# Initialize the SearchEngine
search = SearchEngine()

# Fetch median home value and median household income for each unique ZIP code
zip_data = []
for zip_code in unique_zips:
    result = search.by_zipcode(zip_code)
    if result: # Check if the result is not None
        zip_data.append({
            "zip_code": zip_code,
            "median_home_value": result.median_home_value,
            "median_household_income": result.median_household_income
        })
    else: # Handle the case where the result is None
        zip_data.append({
            "zip_code": zip_code,
            "median_home_value": None,
            "median_household_income": None
        })

# Convert to DataFrame
zip_info_df = pd.DataFrame(zip_data)

# Merge this info back into the original DataFrame based on 'zip_code'
merged_df = pd.merge(valid_zip_df, zip_info_df, how="left", on="zip_code")
```

## 6 Data Manipulation

```
merged_df.head()
```

	crash_datetime	borough	zip_code	latitude	longitude	location
0	2024-06-30 07:05:00	BROOKLYN	11235	40.58106	-73.96744	(40.58106, -73.96744)
1	2024-06-30 20:47:00	MANHATTAN	10021	40.76363	-73.95330	(40.76363, -73.95330)
2	2024-06-30 10:14:00	BROOKLYN	11222	40.73046	-73.95149	(40.73046, -73.95149)
3	2024-06-30 15:52:00	BRONX	10468	40.86685	-73.89597	(40.86685, -73.89597)
4	2024-06-30 16:30:00	BROOKLYN	11226	40.63969	-73.95321	(40.63969, -73.95321)

# 7 Visualization

## 7.1 Data Visualization with Plotnine

This section was written by Julia Mazzola

### 7.1.1 Introduction

Hi! My name is Julia, and I am a Senior double majoring in Economics and Statistical Data Science. I'm excited to show you the power of data visualization with **Plotnine**, a Python library inspired by R's **ggplot2**. Visualization is a crucial tool to effectively communicate your findings to your audience and **Plotnine** is a useful library to use.

### 7.1.2 What is Plotnine?

**Plotnine** uses grammar of graphics to create layered, customizable visualizations. Grammar of graphics is a framework that provides a systematic approach to creating visual representations of data by breaking down the plot into its fundamental components. To understand this better, think about how sentences have grammar, we can layer our graphics to create complex and detailed visualizations.

Components of the layered grammar of graphics:

- **Layer:** used to create the objects on a plot

## 7 Visualization

- **Data:** defines the source of the information to be visualized
  - **Mapping:** defines how the variables are represented in the plot
  - **Statistical transformation (stat):** transforms the data, generally by summarizing the information
  - **Geometric object (geom):** determines the type of plot type (e.g., points, lines, bars)
  - **Position adjustment (position):** adjusts the display of overlapping points to improve clarity
- **Scale:** controls how values are mapped to aesthetic attributes (e.g., color, size)
  - **Coordinate system (coord):** maps the position of objects onto the plane of the plot, and controls how the axes and grid lines are drawn
  - **Faceting (facet):** used to split the data up into subsets of the entire dataset

You can make a wide array of different graphics with **Plotnine**. Some common examples are:

- Scatterplot, `geom_point()`
- Bar Chart `geom_bar()`
- Histogram `geom_histogram()`
- Line Chart `geom_line()`

### 7.1.3 Installing Plotnine

To use **Plotnine** you must install it into your `venv` first. The instructions are as follows:

Type this command into either `conda`, your terminal, `gitbash`, or whatever you use for package install for your `venv`.

For **pip**:

## 7.1 Data Visualization with Plotnine

```
pip install plotnine
```

For **conda**:

```
conda install -c conda-forge plotnine
```

You can import Plotnine without a prefix:

```
from plotnine import *
```

Or with with a prefix to access each component such as:

```
import plotnine as p9
```

This way is generally recommended for larger projects or when collaborating with others for better code maintainability. But for simplicity in this section I will use the first method.

For the examples we will be using NYC open data to visualize motor vehicle crashes from the week of June 30, 2024.

```
import pandas as pd
```

```
nyc_crash = pd.read_feather('nyccrashes_cleaned.feather').dropna(subset=['borough'])
```

### 7.1.4 Scatterplot

Firstly, we will be creating a scatterplot. This can be done with `geom_point()`. Our scatterplot will be displaying Crash Locations based on the longitude and latitude of the crash sites.

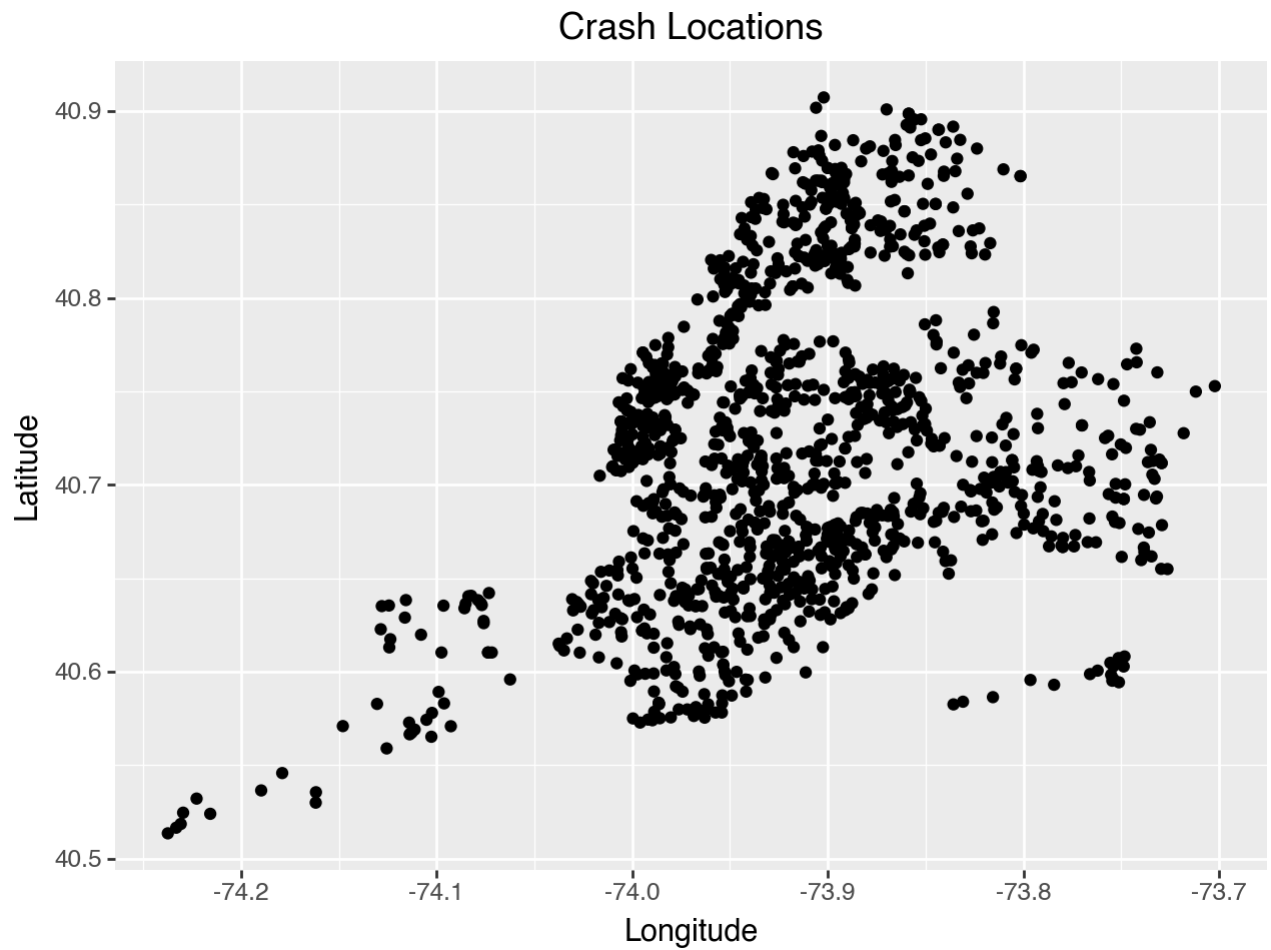
#### Creating a Basic Scatterplot

## 7 Visualization

```
import warnings
warnings.filterwarnings("ignore", category=UserWarning)

(ggplot(nyc_crash, aes(x='longitude', y='latitude')) +
 #specifies graph type
 geom_point() +
 #creates labels for graphic
 labs(title='Crash Locations',
       x='Longitude',
       y='Latitude'))
```





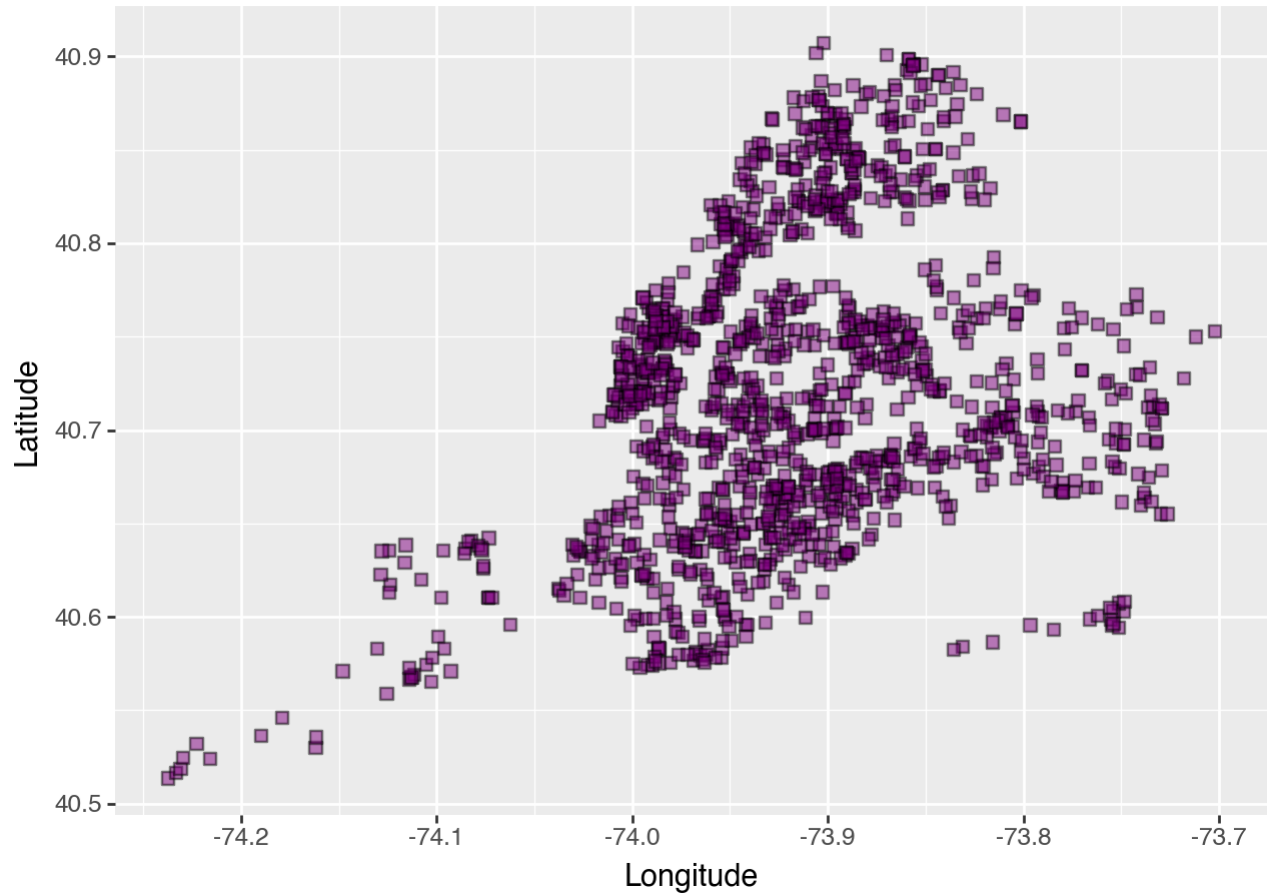
### Customizing a Scatterplot

You can customize your plot further by changing the color, edge color, transparency, size, or shape of your points. This is done in `geom_point()`.

## 7 Visualization

```
(ggplot(nyc_crash, aes(x='longitude', y='latitude')) +  
  geom_point(color = 'black', fill = 'purple',  
             alpha = 0.5, size = 2, shape = 's') +  
  labs(title='Crash Locations by Contributing Factor',  
        x='Longitude',  
        y='Latitude'))
```

### Crash Locations by Contributing Factor



This scatterplot provides a lot of information, yet there are ways we can customize our plot to be more informative for our audience. We can create a scatterplot that differentiates by contributing factor.

#### Changing Shape of Variables

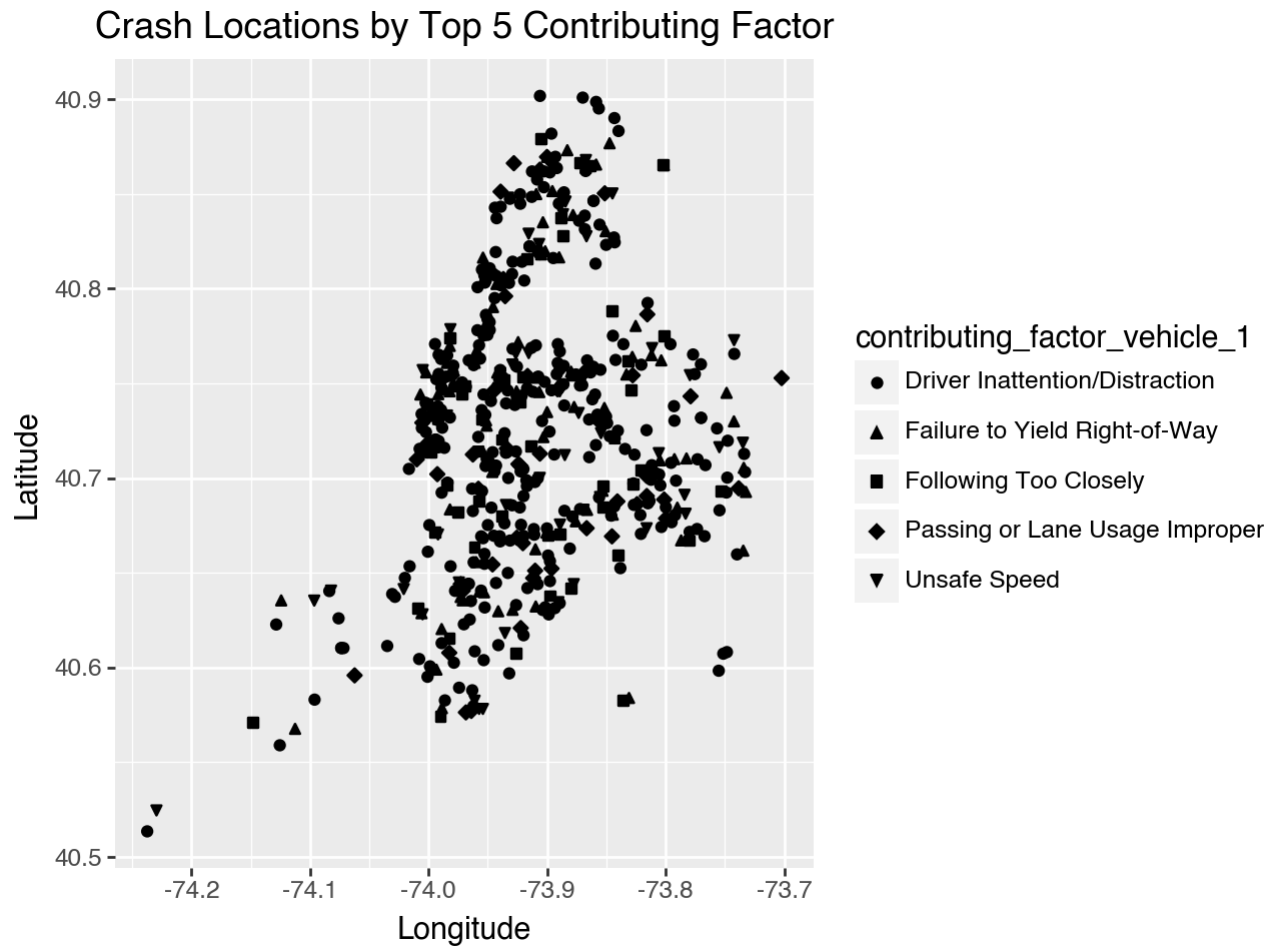
Adding shapes by the contributing factor for vehicle 1:

## 7 Visualization

```
#List of top 5 reasons for the contributing factor since it is difficult to d
factor1 = ["Driver Inattention/Distracted",
          "Failure to Yield Right-of-Way",
          "Following Too Closely",
          "Unsafe Speed",
          "Passing or Lane Usage Improper"]

# Filter the data to only include valid contributing factors
confact = nyc_crash[nyc_crash['contributing_factor_vehicle_1'].isin(factor1)]
```

```
#changes shape of point according to 'contributing_factor_vehicle_1'
(ggplot(confact, aes(x='longitude', y='latitude',
                    shape='contributing_factor_vehicle_1')) +
 geom_point() +
 labs(title='Crash Locations by Top 5 Contributing Factor',
      x='Longitude',
      y='Latitude'))
```

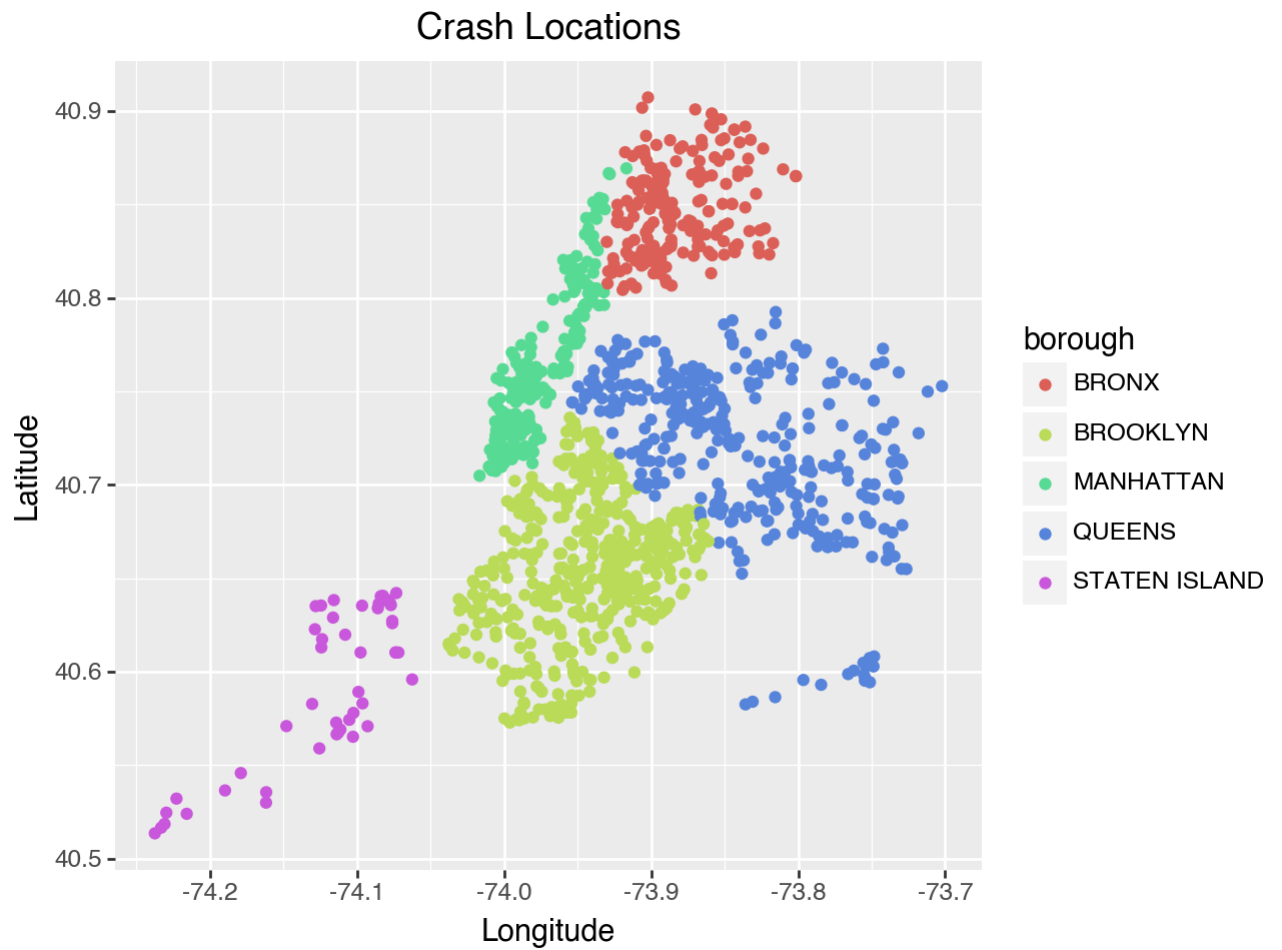


### Changing Color of Variables

To add color coordination to your plot in Plotnine, you can specify the variable you want to use for coloring by including `color='variable'` within the `aes()` function. This allows you to visually distinguish different categories in your dataset.

## 7 Visualization

```
#color= changes color according to 'borough'
(ggplot(nyc_crash, aes(x='longitude', y='latitude', color = 'borough')) +
  geom_point() +
  labs(title='Crash Locations',
        x='Longitude',
        y='Latitude'))
```

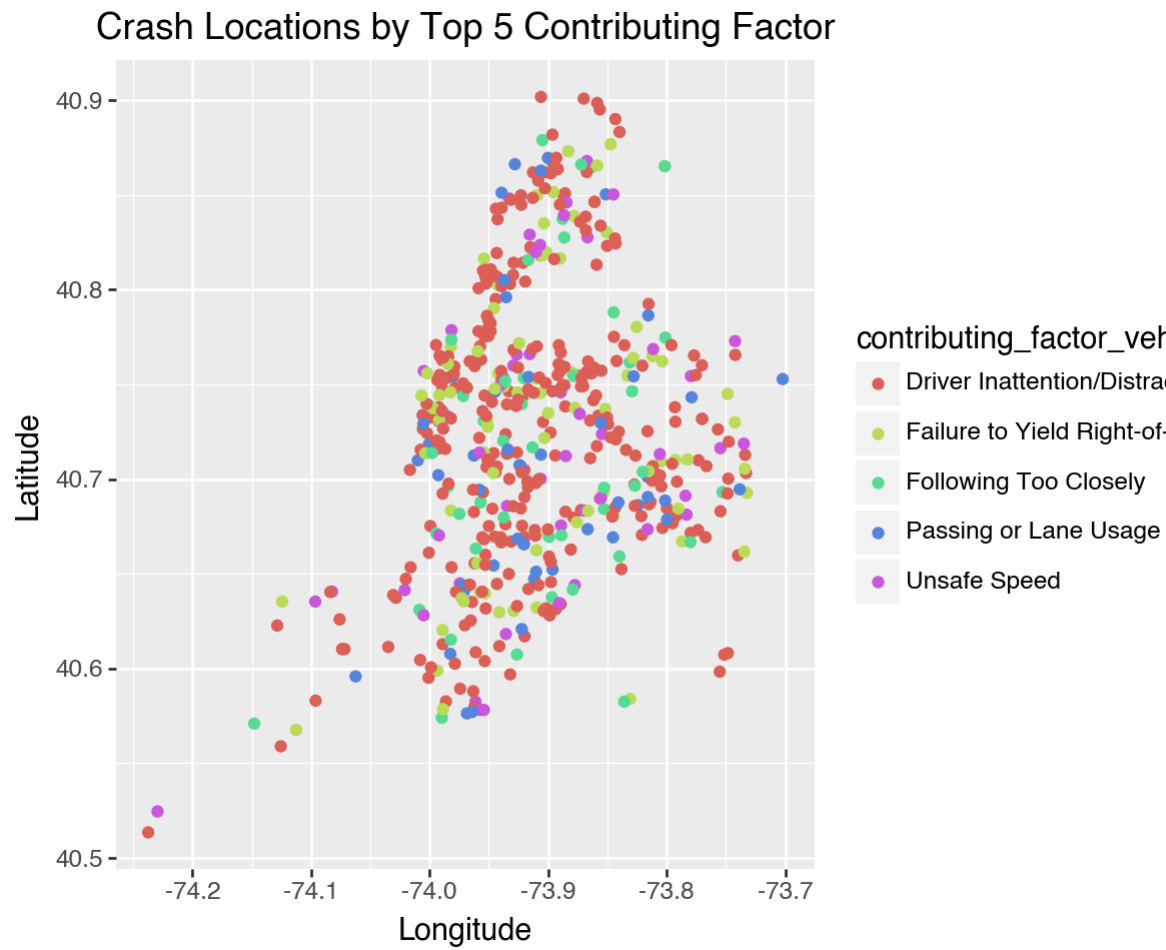


As you can see each borough has its own color and the audience can easily see which borough the crash occurred in.

```
#color= changes color according to 'contributing_factor_vehicle_1'
(ggplot(confact, aes(x='longitude', y='latitude',
  color = 'contributing_factor_vehicle_1')) +
```

## 7 Visualization

```
geom_point() +  
labs(title='Crash Locations by Top 5 Contributing Factor',  
      x='Longitude',  
      y='Latitude'))
```



This graph uses color to distinguish what contributing factor caused the



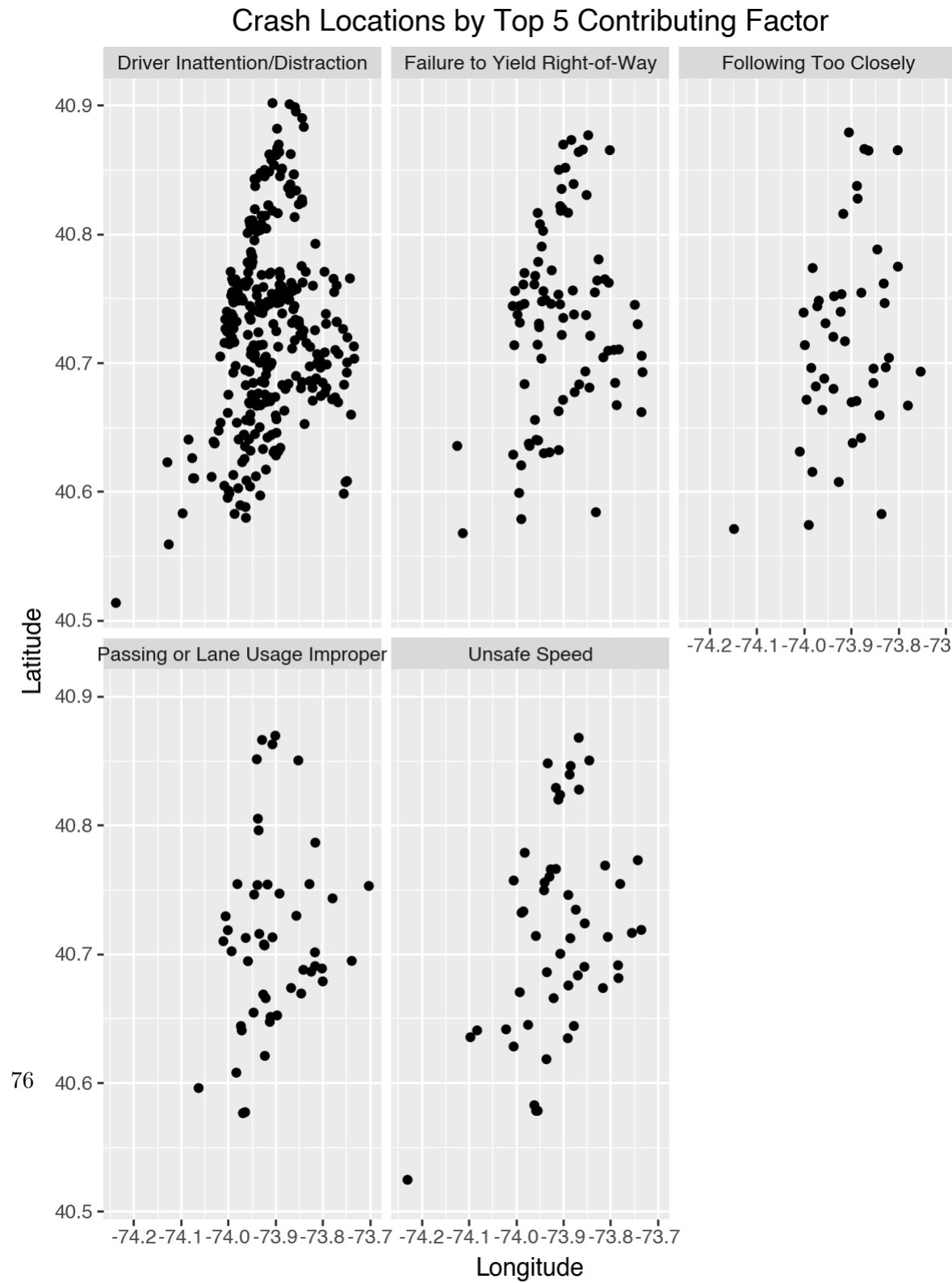
crash.

### Scatterplots per Facet

To organize your data in a way that enhances interpretability, you can utilize `facet_grid()` or `facet_wrap()`. This approach allows for the creation of separate plots based on categorical variables, making it easier to identify trends and patterns.

Scatterplot of Crash Locations by Contributing Factor with `facet_wrap()`:

```
(ggplot(confact, aes(x='longitude', y='latitude')) +  
  geom_point() +  
  #create separate plots for each contributing factor  
  facet_wrap('contributing_factor_vehicle_1') +  
  labs(title='Crash Locations by Top 5 Contributing Factor',  
        x='Longitude',  
        y='Latitude') +  
  #changes the size of the graphic  
  theme(figure_size=(6, 8)))
```

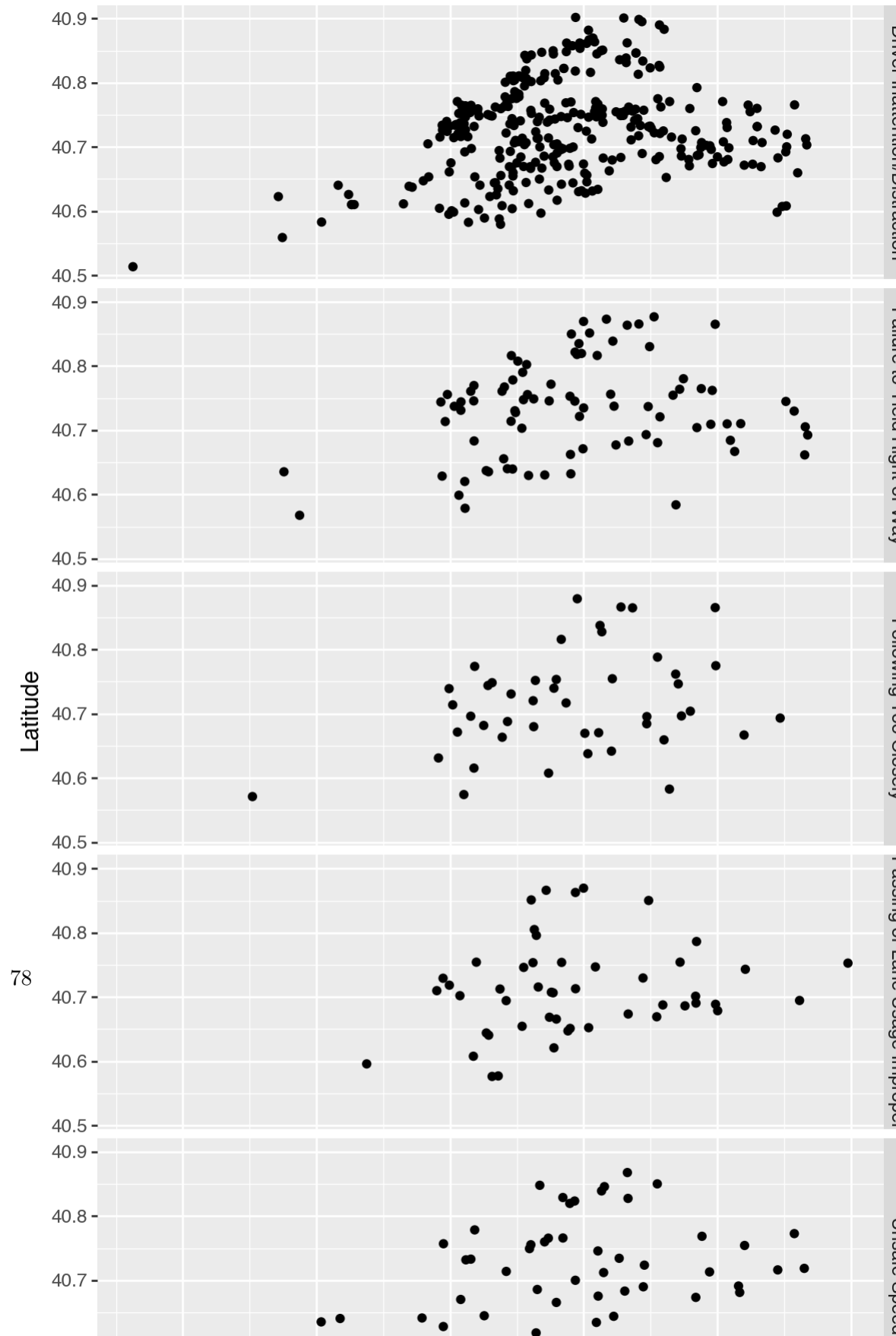


## 7.1 Data Visualization with Plotnine

Scatterplot of Crash Locations by Contributing Factor with `facet_grid()`:

```
(ggplot(confact, aes(x='longitude', y='latitude')) +  
  geom_point() +  
  #create separate plots for each contributing factor  
  facet_grid('contributing_factor_vehicle_1') +  
  labs(title='Crash Locations by Top 5 Contributing Factor',  
        x='Longitude',  
        y='Latitude') +  
  #changes the size of the graphic  
  theme(figure_size=(6, 10)))
```

Crash Locations by Top 5 Contributing Factor

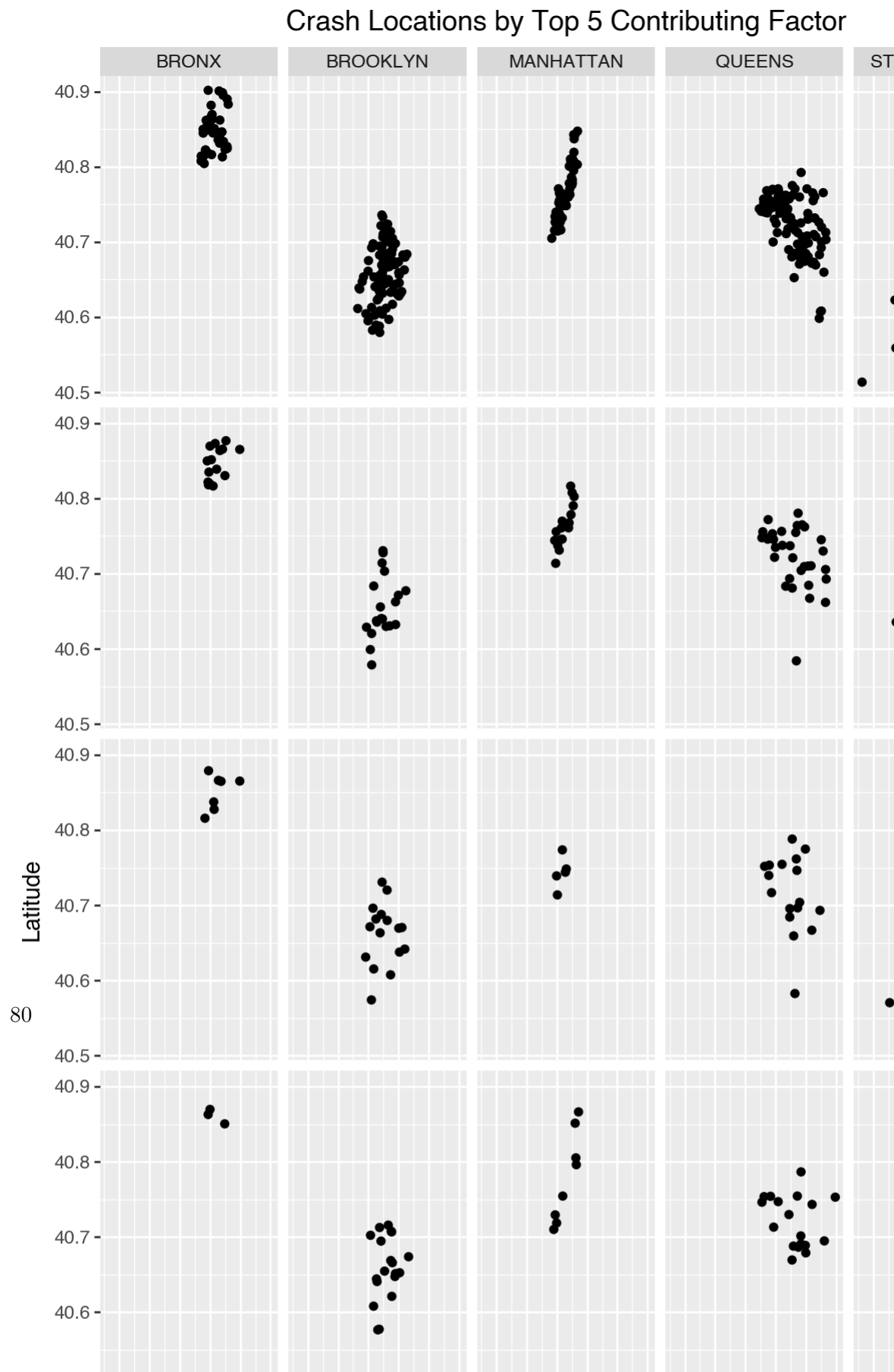


## 7.1 Data Visualization with Plotnine

Scatterplot of Crash Locations by Contributing Factor and Borough:

```
(ggplot(confact, aes(x='longitude', y='latitude')) +  
  geom_point() +  
  #create a grid of subplots based on the values of two categorical variables  
  facet_grid('contributing_factor_vehicle_1 ~ borough') +  
  labs(title='Crash Locations by Top 5 Contributing Factor',  
        x='Longitude',  
        y='Latitude') +  
  #changes angle of text and size of the graphic  
  theme(  
    axis_text_x=element_text(angle=90),  
    figure_size=(7, 12)))
```

7 Visualization



## Adding linear regression line to scatterplot

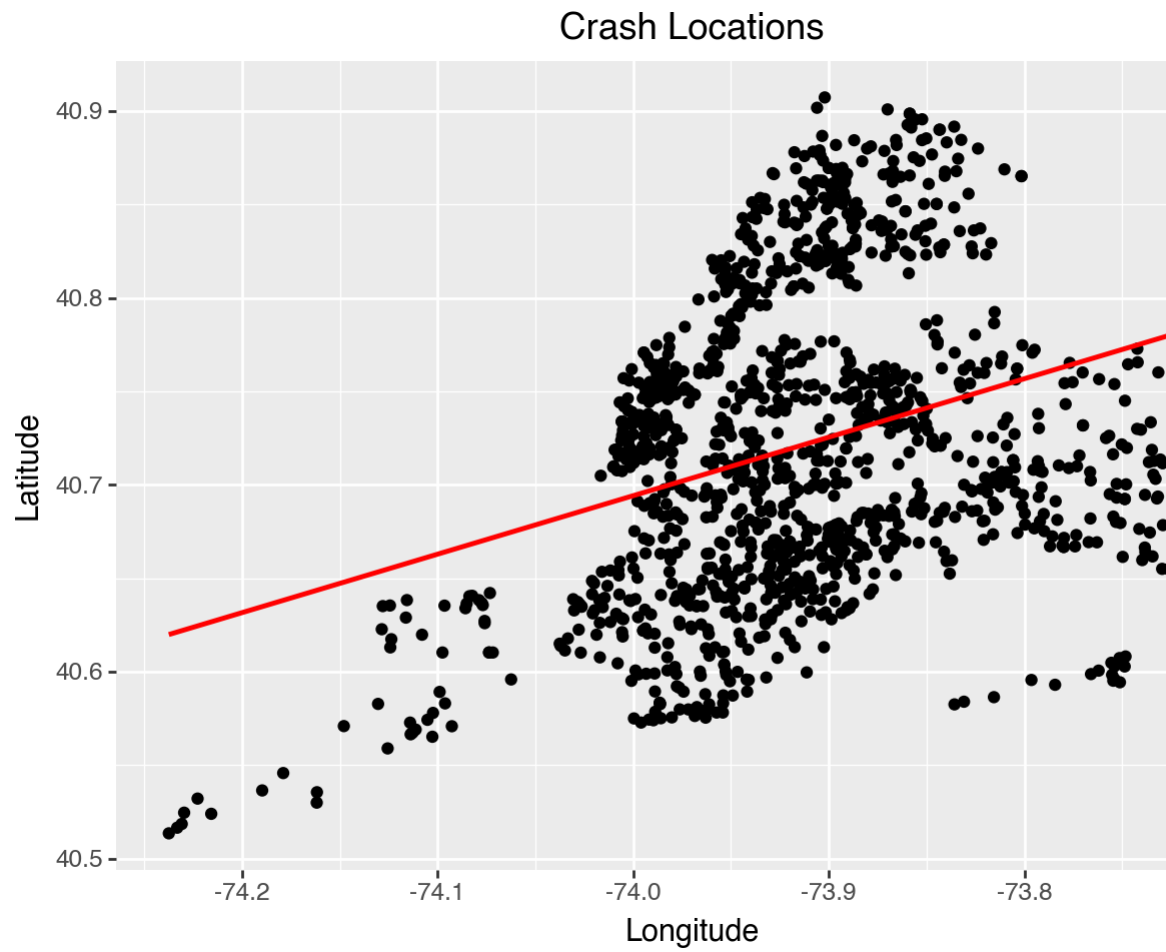
Fitting a regression line to your plot can be really helpful to visualize trends to your data easier. To add a linear regression line to your scatterplot, you would include the following line of code:

```
geom_smooth(method='lm', se=False, color=' ')
```

```
<plotnine.geoms.geom_smooth.geom_smooth at 0x114b7eea0>
```

Although this data is not one that makes sense to fit a regression line, the scatterplot with it is as follows:

```
(ggplot(nyc_crash, aes(x='longitude', y='latitude')) +  
  geom_point() +  
  #adds a linear regression line  
  geom_smooth(method='lm', se=False, color='red') +  
  labs(title='Crash Locations',  
        x='Longitude',  
        y='Latitude'))
```



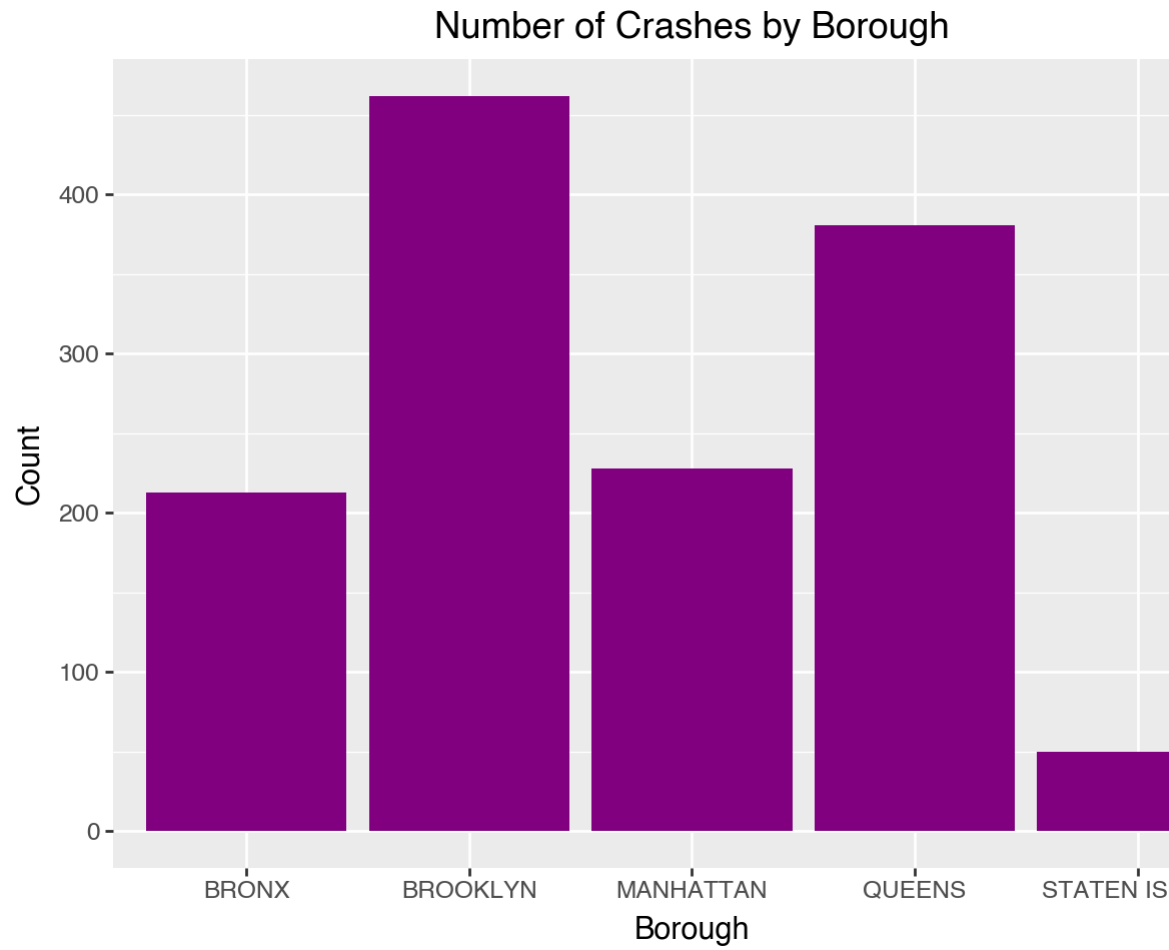
### 7.1.5 Bar Chart

Another common use for displaying data is a bar chart. You can create one with `geom_bar()`. We will start with a simple chart of crashes by borough.



## Creating a Basic Bar Chart

```
(ggplot(nyc_crash, aes(x='borough')) # Use 'borough' for the x-axis
+ geom_bar(fill='purple')
+ labs(title='Number of Crashes by Borough',
       x='Borough',
       y='Count'))
```



### Customizing your Bar Chart

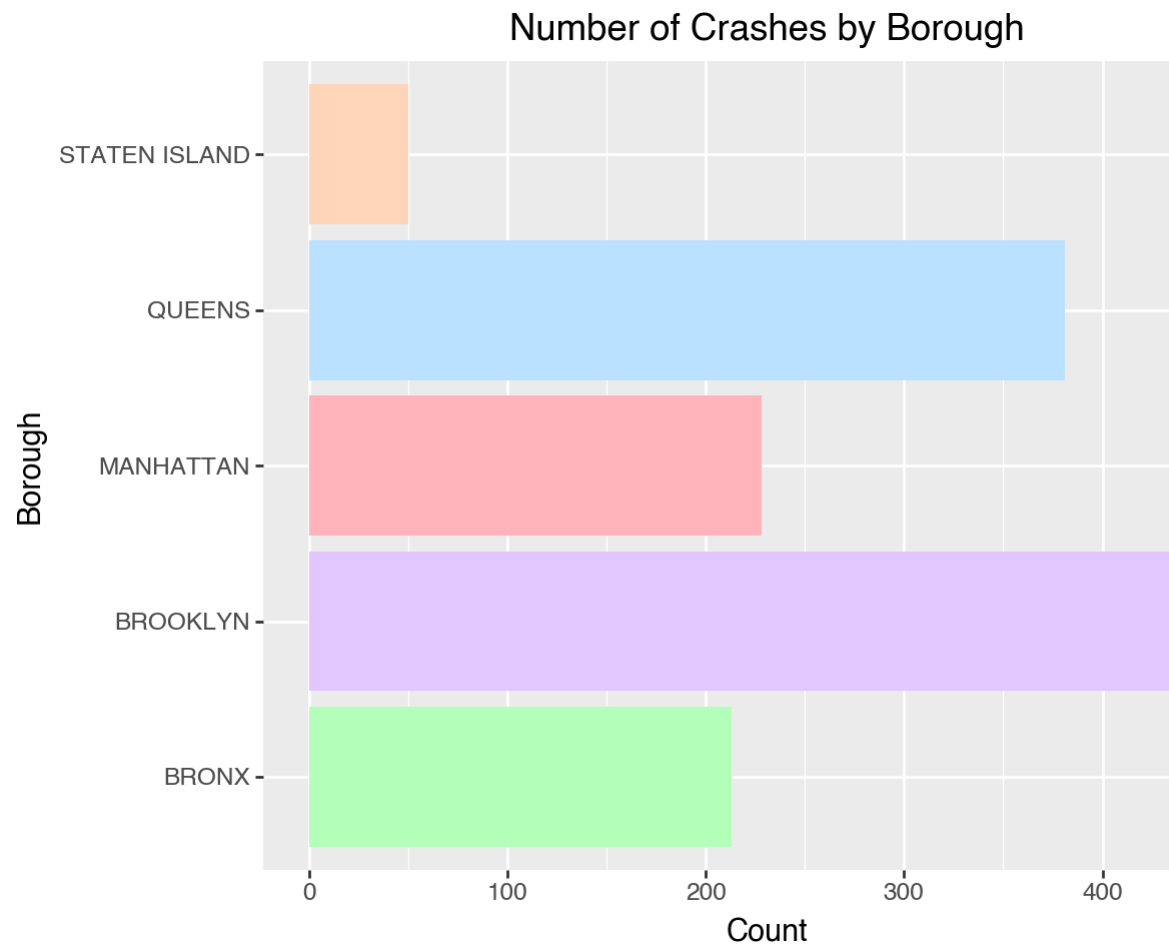
You can change up your bar chart a couple of different ways. You can handpick colors you want, designate it to variables, and flip orientation, etc:

## 7.1 Data Visualization with Plotnine

```
#designate your preferred colors (pastel color codes)
colors = ['#B3FFBA', '#E1C6FF', '#FFB3BA', '#BAE1FF', '#FFD5BA']

#adding fill= changes the color of bar according to variable
(ggplot(nyc_crash, aes(x='borough', fill = 'borough'))
#assigns your preferred colors
+ geom_bar(fill = colors)
#flips orientation of the chart
+ coord_flip()
+ labs(title='Number of Crashes by Borough',
        x='Borough',
        y='Count'))
```

## 7 Visualization



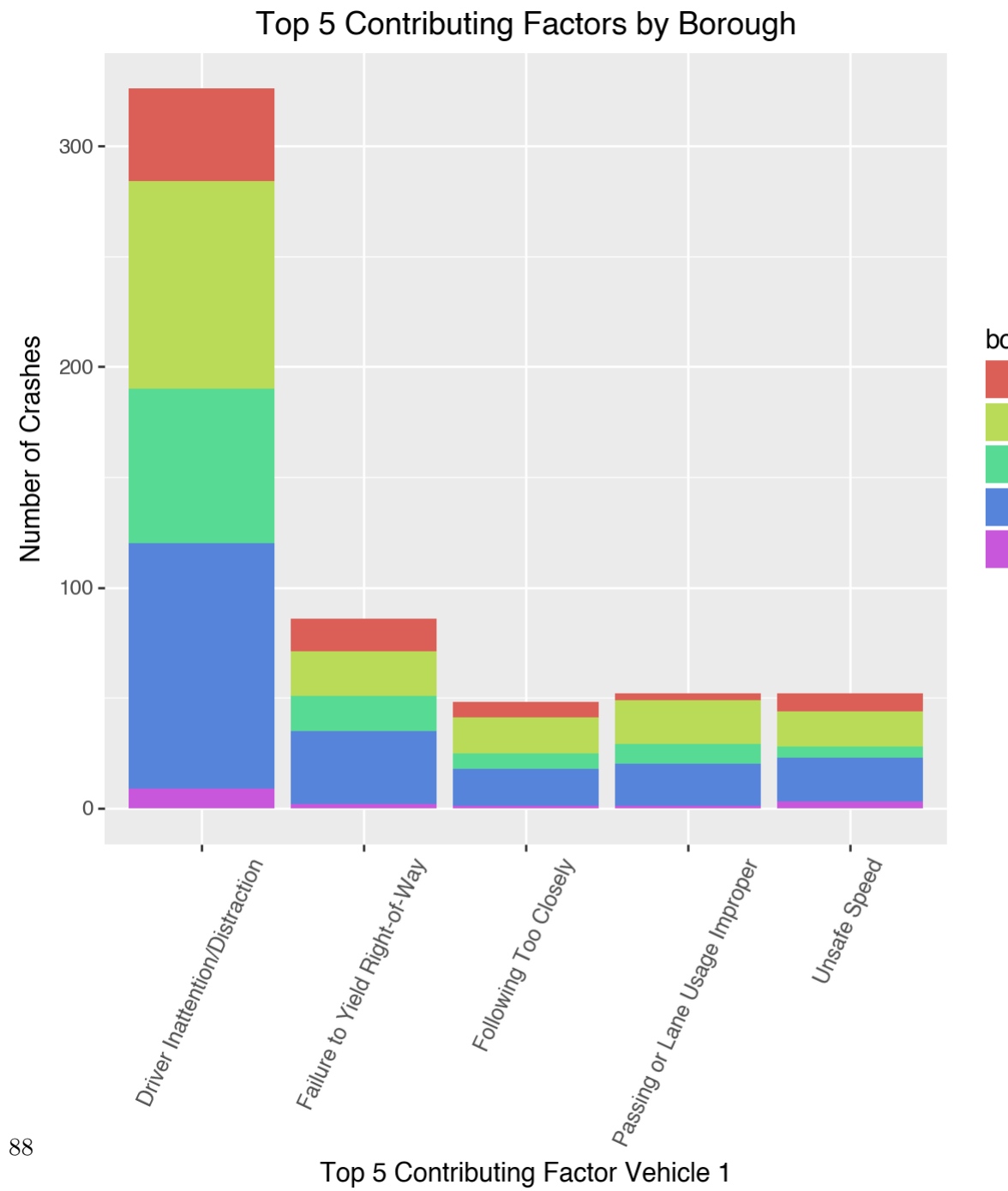
### Multivariable Bar Chart

You can also split up a bar chart to make it visually easier to understand.

```
#using 'confact' dataset again for better visualization
(ggplot(confact, aes(x='contributing_factor_vehicle_1', fill='borough'))
 + geom_bar())
```

## 7.1 Data Visualization with Plotnine

```
+ ggtitle('Top 5 Contributing Factors by Borough')
+ xlab("Top 5 Contributing Factor Vehicle 1")
+ ylab("Number of Crashes")
#creates smaller text
#rotates x-axis text for readability
#creates a bigger image
+ theme(axis_text_x=element_text(size=9, angle=65), figure_size= (7,7))
```

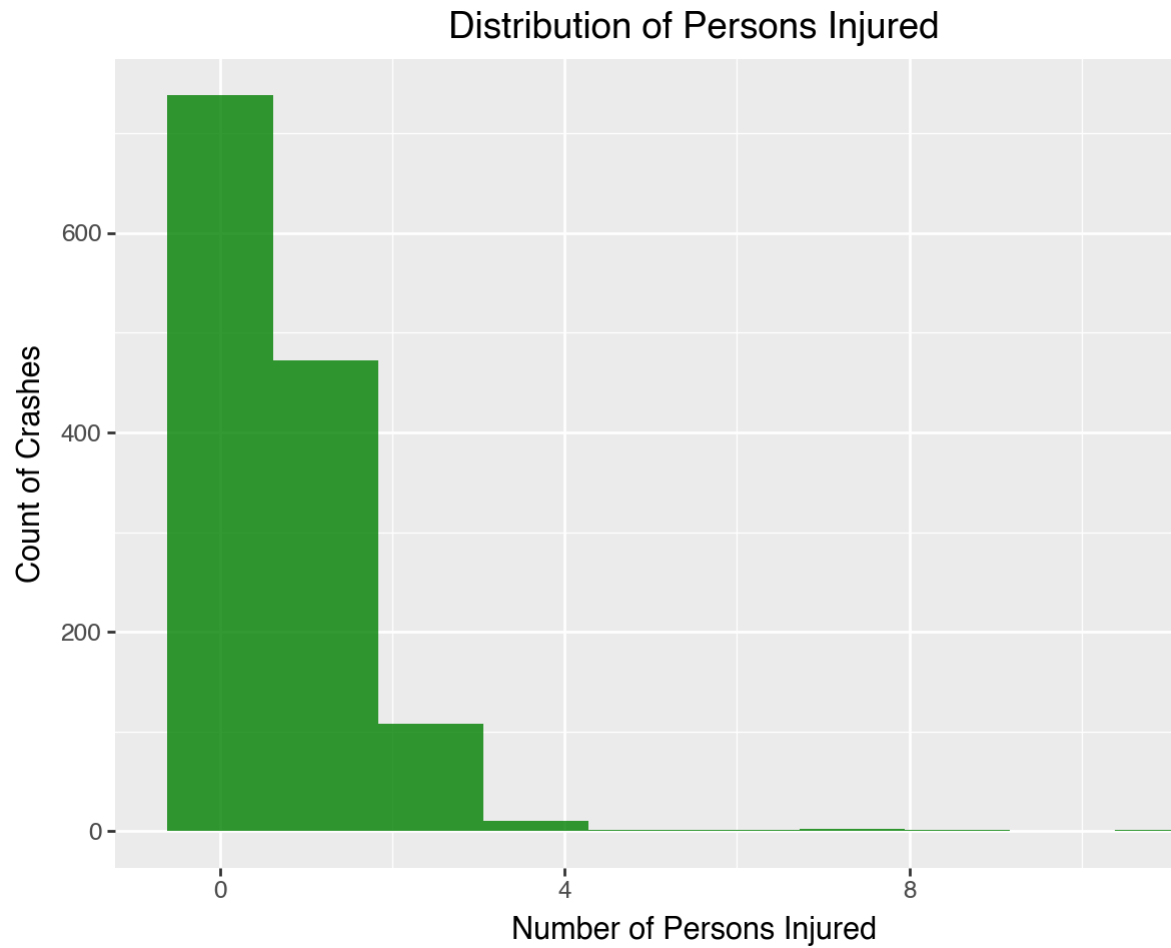


### 7.1.6 Histogram

Another useful way to display data is a histogram. You can create one with `geom_histogram()`. Using a histogram is very useful when displaying continuous data.

#### Basic Histogram

```
(ggplot(nyc_crash, aes(x='number_of_persons_injured')) +  
  #bins= sets the amount of bars in your histogram  
  geom_histogram(bins=10, alpha=0.8, fill='green') +  
  labs(title='Distribution of Persons Injured',  
        x='Number of Persons Injured',  
        y='Count of Crashes'))
```



With a histogram it is very easy to understand trends for a dataset and you can see that our NYC crash data is positively skewed.

### Multivariable Histogram

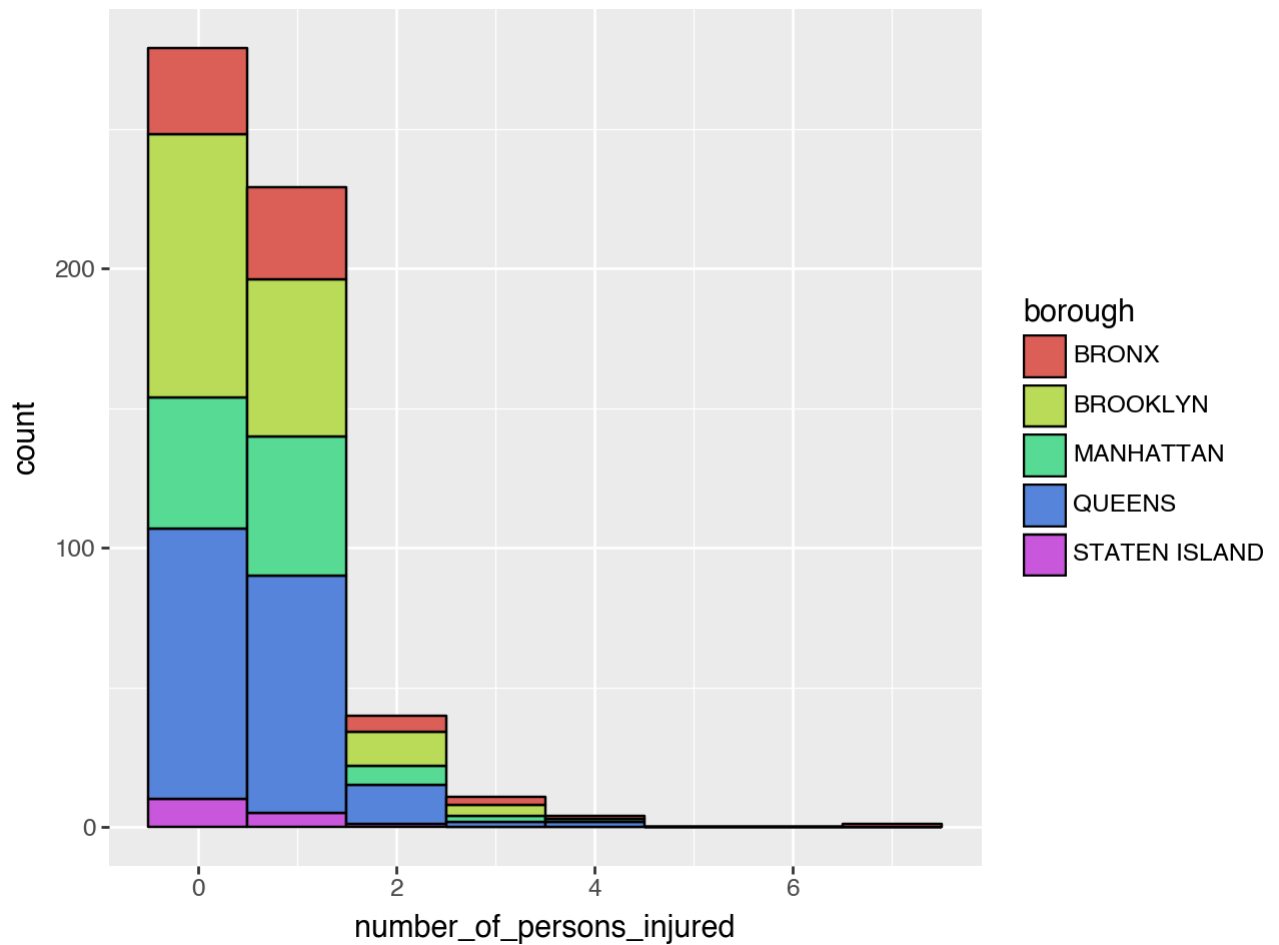
Similar to bar charts, we can make Histograms that display more than one variable. We can also add some more customizations such as adding an



## 7.1 Data Visualization with Plotnine

outline color for our sections, adding width of the bins and angle of .

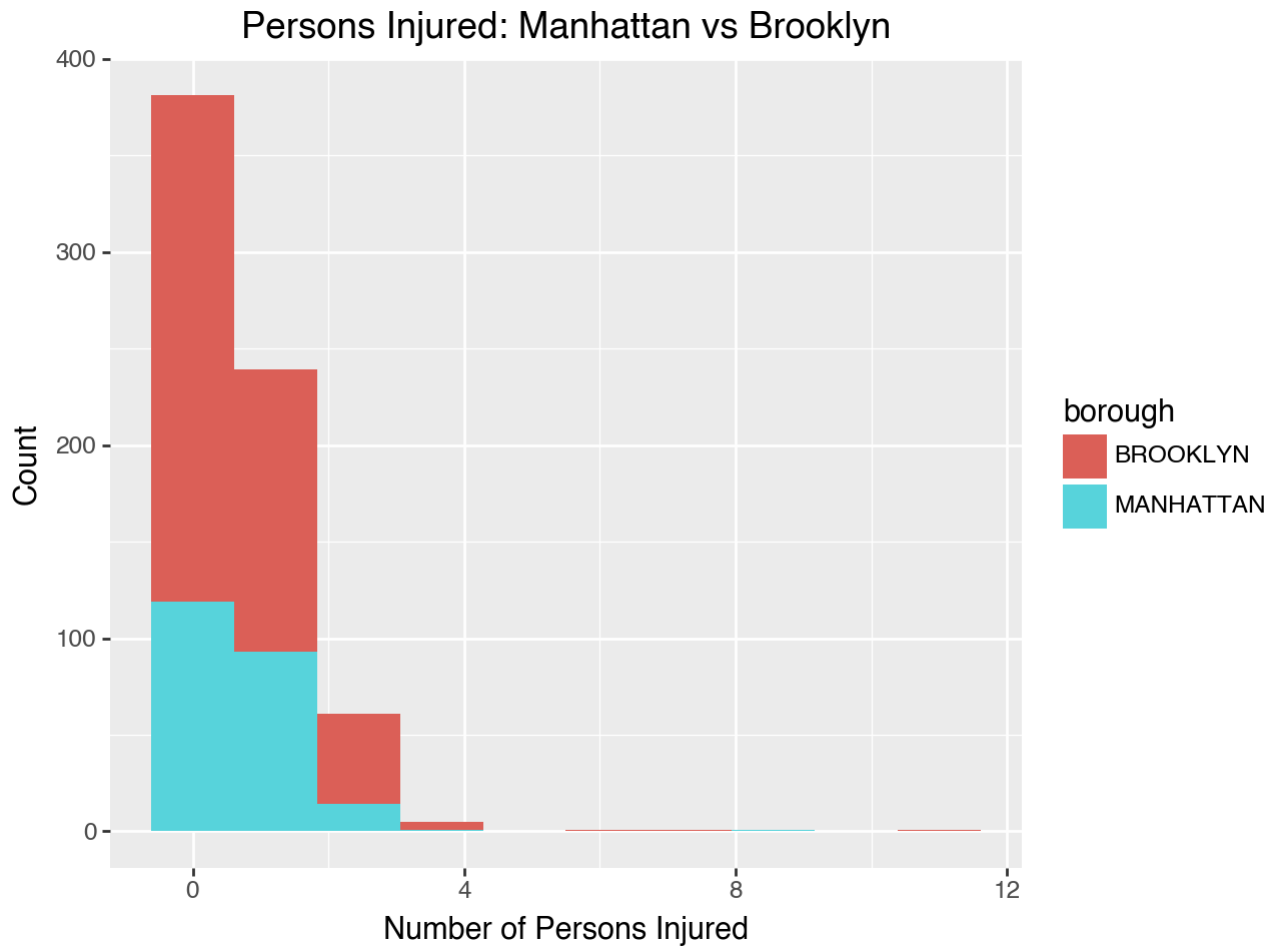
```
(ggplot(confact, aes(x='number_of_persons_injured', fill = 'borough'))  
  #changing width of the bins and adding outline color with color= to make it easier to re  
  + geom_histogram(binwidth=1, color = 'black'))
```



## Overlapping Histogram

Histograms can also be useful when comparing multiple categories. Here we are comparing Manhattan and Brooklyn's number of persons injured with an overlapping histogram.

```
(ggplot(nyc_crash[nyc_crash['borough'].isin(['MANHATTAN', 'BROOKLYN'])], aes(
  geom_histogram(bins=10) +
  labs(title='Persons Injured: Manhattan vs Brooklyn',
        x='Number of Persons Injured',
        y='Count'))
```



### 7.1.7 Line Chart

Line charts are great for time-series data and can be created with `geom_line()`. This type of chart is particularly useful for identifying patterns, fluctuations, and trends, making it easier to understand how

## 7 Visualization

a variable changes over a specified period. We will create one analyzing Number of Crashes by Hour.

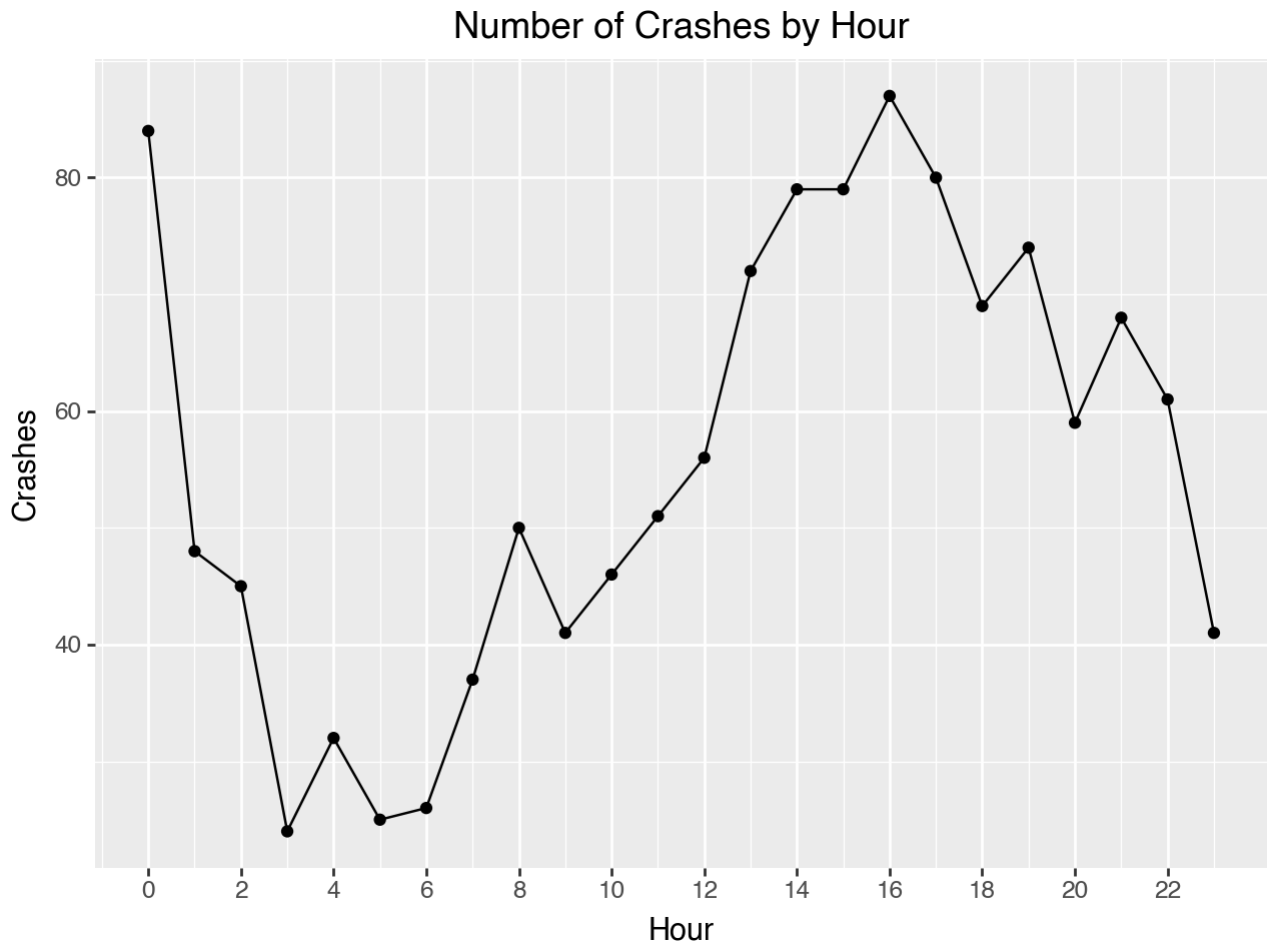
### Basic Line Chart

```
nyc_crash['crash_datetime'] = pd.to_datetime(nyc_crash['crash_datetime'])

# Extract hour
nyc_crash['crash_hour'] = nyc_crash['crash_datetime'].dt.hour

# Count crashes per hour
crash_counts = nyc_crash.groupby(['crash_hour']).size().reset_index(name='crash_count')

#plot crashes by hour
(ggplot(crash_counts, aes(x='crash_hour', y='crash_count')) +
 #creates the line chart
 geom_line() +
 #adds points for better visibility
 geom_point() +
 labs(title='Number of Crashes by Hour',
       x='Hour',
       y='Crashes') +
 #formats the x-axis to display ticks by every 2 hours
 scale_x_continuous(breaks=range(0, 24, 2)))
```



This example is excellent for understanding the grammar of graphics. As you can see, we use `geom_line()` to create the line chart, while also adding `geom_point()`, which is typically used for scatterplots, to make the figure clearer by layering additional details.”

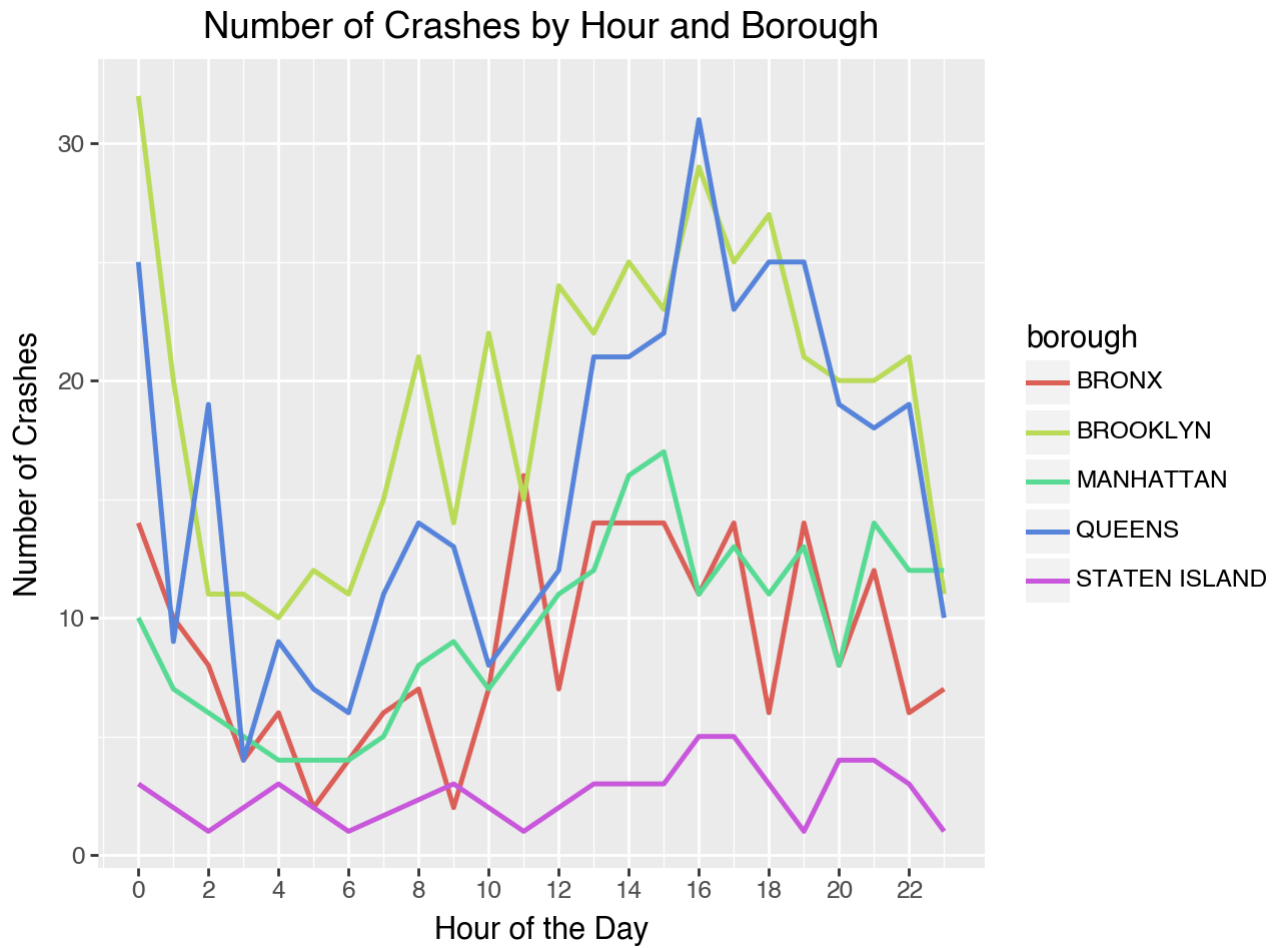
### Multivariable Line Chart

## 7 Visualization

Similarly to the other figures you can create a line chart with multiple variables. Now we will create a chart with number of crashes by borough.

```
#setting crash counts to also include borough
crash_counts = nyc_crash.groupby(['crash_hour', 'borough']).size().reset_index()

#plots crashes by hour with different lines for each borough
(ggplot(crash_counts, aes(x='crash_hour', y='crash_count', color='borough'))
 #size= changes the thinkness of the lines
 geom_line(size=1) +
 labs(title='Number of Crashes by Hour and Borough',
      x='Hour of the Day',
      y='Number of Crashes') +
 scale_x_continuous(breaks=range(0, 24, 2)))
```



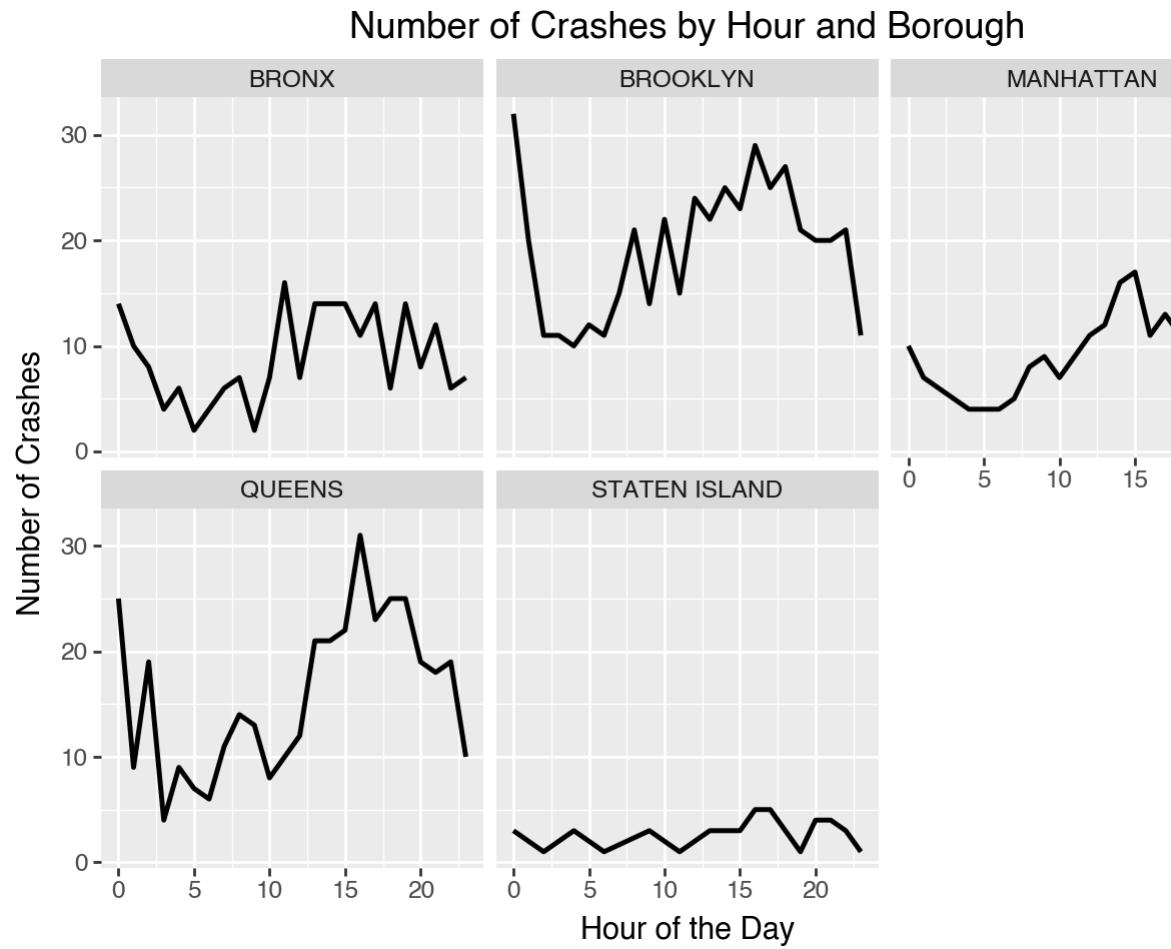
### Line Chart per Facet

You can use plot each variable by on separate panels with `facet_wrap()`.

```
(ggplot(crash_counts, aes(x='crash_hour', y='crash_count')) +  
  geom_line(size=1) +  
  #breaks the figure up by borough
```

## 7 Visualization

```
facet_wrap("borough") +  
labs(title='Number of Crashes by Hour and Borough',  
     x='Hour of the Day',  
     y='Number of Crashes'))
```





### 7.1.8 Conclusion

Plotnine is a very powerful tool to make impactful and detailed graphics. The flexibility of its grammar of graphics approach means there are endless ways to modify, enhance, and be creative with your plots. You can layer geoms, adjust aesthetics, and apply scales, facets, and themes.

#### Basic things to remember:

#### Creating Specific Plots

- Scatterplot, `geom_point()`
- Boxplot `geom_box()`
- Histogram `geom_histogram()`
- Line Chart `geom_line()`
- Bar Chart `geom_bar()`
- Density Plot `geom_density()`

#### Formatting and Customizing Your Figure

- `fill`: to change the color of the data
- `color`: to change the color of the borders
- `alpha`: to change the transparency
- `bins`: to change the number of bins
- `figure_size`: to change size of graphic
- `geom_smooth`: to add a smoothed line
- `facet`: plot each group on a separate panel
  - `facet_wrap()`: creates a series of plots arranged in a grid, wrapping into new rows or columns as needed
  - `facet_grid()`: allows you to create a grid layout based on two categorical variables, organizing plots in a matrix format
- `theme`: change overall theme

## 7 Visualization

There are many other features and customizations you can do with **Plotnine**. For more information on how to leverage the full potential of this package for your data visualization needs check out Plotnine’s Graph Gallery.

Happy plotting!

### 7.1.9 Sources

Sarker, D. (2018). A comprehensive guide to the grammar of graphics for effective visualization of multi-dimensional data. Towards Data Science. <https://towardsdatascience.com/a-comprehensive-guide-to-the-grammar-of-graphics-for-effective-visualization-of-multi-dimensional-1f92b4ed4149>

Python Graph Gallery. (2024). Plotnine: ggplot in python. Python Graph Gallery. <https://python-graph-gallery.com/plotnine/>

## 8 Statistical Tests and Models

### 8.1 Tests for Exploratory Data Analysis

A collection of functions are available from `scipy.stats`.

- Comparing the locations of two samples
  - `ttest_ind`: t-test for two independent samples
  - `ttest_rel`: t-test for paired samples
  - `ranksums`: Wilcoxon rank-sum test for two independent samples
  - `wilcoxon`: Wilcoxon signed-rank test for paired samples
- Comparing the locations of multiple samples
  - `f_oneway`: one-way ANOVA
  - `kruskal`: Kruskal-Wallis H-test
- Tests for associations in contingency tables
  - `chi2_contingency`: Chi-square test of independence of variables
  - `fisher_exact`: Fisher exact test on a 2x2 contingency table
- Goodness of fit
  - `goodness_of_fit`: distribution could contain unspecified parameters
  - `anderson`: Anderson-Darling test
  - `kstest`: Kolmogorov-Smirnov test

## 8 Statistical Tests and Models

- `chisquare`: one-way chi-square test
- `normaltest`: test for normality

Since R has a richer collections of statistical functions, we can call R function from Python with `rpy2`. See, for example, a blog on this subject.

For example, `fisher_exact` can only handle 2x2 contingency tables. For contingency tables larger than 2x2, we can call `fisher.test()` from R through `rpy2`. See this StackOverflow post. Note that the `.` in function names and arguments are replaced with `_`.

```
import pandas as pd
import numpy as np
import rpy2.robjects.numpy2ri
from rpy2.robjects.packages import importr
rpy2.robjects.numpy2ri.activate()

stats = importr('stats')

w0630 = pd.read_feather("data/nyccrashes_cleaned.feather")
w0630["injury"] = np.where(w0630["number_of_persons_injured"] > 0, 1, 0)
m = pd.crosstab(w0630["injury"], w0630["borough"])
print(m)

res = stats.fisher_test(m.to_numpy(), simulate_p_value = True)
print(res)
```

Loading custom .Rprofile	borough	BRONX	BROOKLYN	MANHATTAN	QUEENS	STATEN I
injury						
0	149	345	164	249	65	
1	129	266	127	227	28	

Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

```
data:  structure(c(149L, 129L, 345L, 266L, 164L, 127L, 249L, 227L, 65L, 28L), dim = c(2L, 5L))
p-value = 0.03848
alternative hypothesis: two.sided
```

## 8.2 Statistical Modeling

Statistical modeling is a cornerstone of data science, offering tools to understand complex relationships within data and to make predictions. Python, with its rich ecosystem for data analysis, features the `statsmodels` package— a comprehensive library designed for statistical modeling, tests, and data exploration. `statsmodels` stands out for its focus on classical statistical models and compatibility with the Python scientific stack (`numpy`, `scipy`, `pandas`).

### 8.2.1 Installation of `statsmodels`

To start with statistical modeling, ensure `statsmodels` is installed:

Using pip:

```
pip install statsmodels
```

### 8.2.2 Linear Model

Let's simulate some data for illustrations.

## 8 Statistical Tests and Models

```
import numpy as np

nobs = 200
ncov = 5
np.random.seed(123)
x = np.random.random((nobs, ncov)) # Uniform over [0, 1)
beta = np.repeat(1, ncov)
y = 2 + np.dot(x, beta) + np.random.normal(size = nobs)
```

Check the shape of y:

```
y.shape
```

```
(200,)
```

Check the shape of x:

```
x.shape
```

```
(200, 5)
```

That is, the true linear regression model is  $[y = 2 + x\_1 + x\_2 + x\_3 + x\_4 + x\_5 + .]$

A regression model for the observed data can be fitted as

```
import statsmodels.api as sma
xmat = sma.add_constant(x)
mymod = sma.OLS(y, xmat)
myfit = mymod.fit()
myfit.summary()
```

## 8.2 Statistical Modeling

<b>Dep. Variable:</b>	y	<b>R-squared:</b>	0.309
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.292
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	17.38
<b>Date:</b>	Mon, 07 Oct 2024	<b>Prob (F-statistic):</b>	3.31e-14
<b>Time:</b>	13:17:03	<b>Log-Likelihood:</b>	-272.91
<b>No. Observations:</b>	200	<b>AIC:</b>	557.8
<b>Df Residuals:</b>	194	<b>BIC:</b>	577.6
<b>Df Model:</b>	5		
<b>Covariance Type:</b>	nonrobust		

	coef	std err	t	P>  t	[0.025	0.975]
<b>const</b>	1.8754	0.282	6.656	0.000	1.320	2.431
<b>x1</b>	1.1703	0.248	4.723	0.000	0.682	1.659
<b>x2</b>	0.8988	0.235	3.825	0.000	0.435	1.362
<b>x3</b>	0.9784	0.238	4.114	0.000	0.509	1.448
<b>x4</b>	1.3418	0.250	5.367	0.000	0.849	1.835
<b>x5</b>	0.6027	0.239	2.519	0.013	0.131	1.075

<b>Omnibus:</b>	0.810	<b>Durbin-Watson:</b>	1.978
<b>Prob(Omnibus):</b>	0.667	<b>Jarque-Bera (JB):</b>	0.903
<b>Skew:</b>	-0.144	<b>Prob(JB):</b>	0.637
<b>Kurtosis:</b>	2.839	<b>Cond. No.</b>	8.31

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Questions to review:

- How are the regression coefficients interpreted? Intercept?
- Why does it make sense to center the covariates?

Now we form a data frame with the variables

## 8 Statistical Tests and Models

```
import pandas as pd
df = np.concatenate((y.reshape((nobs, 1)), x), axis = 1)
df = pd.DataFrame(data = df,
                  columns = ["y"] + ["x" + str(i) for i in range(1,
ncov + 1)])
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 6 columns):
#   Column  Non-Null Count  Dtype
---  ------  -
0    y      200 non-null     float64
1   x1      200 non-null     float64
2   x2      200 non-null     float64
3   x3      200 non-null     float64
4   x4      200 non-null     float64
5   x5      200 non-null     float64
dtypes: float64(6)
memory usage: 9.5 KB
```

Let's use a formula to specify the regression model as in R, and fit a robust linear model (`rlm`) instead of OLS. Note that the model specification and the function interface is similar to R.

```
import statsmodels.formula.api as smf
mymod = smf.rlm(formula = "y ~ x1 + x2 + x3 + x4 + x5", data = df)
myfit = mymod.fit()
myfit.summary()
```



## 8.2 Statistical Modeling

<b>Dep. Variable:</b>	y	<b>No. Observations:</b>	200
<b>Model:</b>	RLM	<b>Df Residuals:</b>	194
<b>Method:</b>	IRLS	<b>Df Model:</b>	5
<b>Norm:</b>	HuberT		
<b>Scale Est.:</b>	mad		
<b>Cov Type:</b>	H1		
<b>Date:</b>	Mon, 07 Oct 2024		
<b>Time:</b>	13:17:03		
<b>No. Iterations:</b>	16		

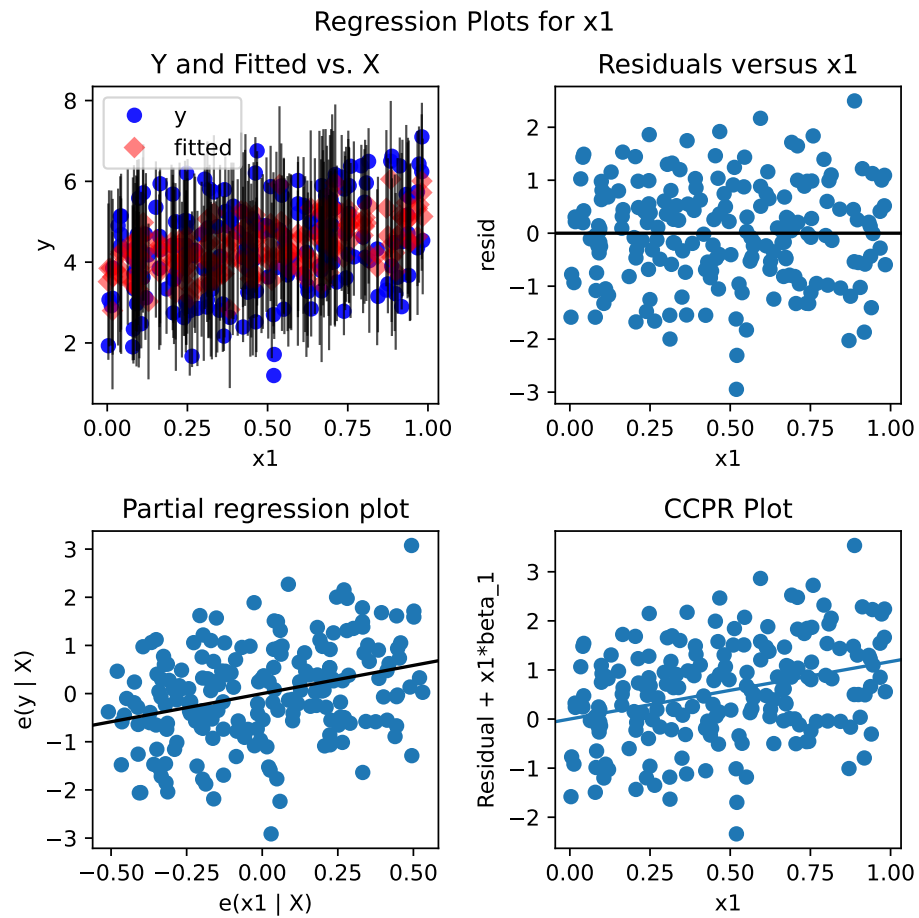
	coef	std err	z	P> z	[0.025	0.975]
<b>Intercept</b>	1.8353	0.294	6.246	0.000	1.259	2.411
<b>x1</b>	1.1254	0.258	4.355	0.000	0.619	1.632
<b>x2</b>	0.9664	0.245	3.944	0.000	0.486	1.447
<b>x3</b>	0.9995	0.248	4.029	0.000	0.513	1.486
<b>x4</b>	1.3275	0.261	5.091	0.000	0.816	1.839
<b>x5</b>	0.6768	0.250	2.712	0.007	0.188	1.166

If the model instance has been used for another fit with different fit parameters, then the fit options might not be the correct ones anymore .

For model diagnostics, one can check residual plots.

```
import matplotlib.pyplot as plt

myOlsFit = smf.ols(formula = "y ~ x1 + x2 + x3 + x4 + x5", data = df).fit()
fig = plt.figure(figsize = (6, 6))
## residual versus x1; can do the same for other covariates
fig = sma.graphics.plot_regress_exog(myOlsFit, 'x1', fig=fig)
```



See more on residual diagnostics and specification tests.

### 8.2.3 Generalized Linear Regression

A linear regression model cannot be applied to presence/absence or count data. Generalized Linear Models (GLM) extend the classical linear regres-

sion to accommodate such response variables, that follow distributions other than the normal distribution. GLMs consist of three main components:

- **Random Component:** This specifies the distribution of the response variable  $Y$ . It is assumed to be from the exponential family of distributions, such as Binomial for binary data and Poisson for count data.
- **Systematic Component:** This consists of the linear predictor, a linear combination of unknown parameters and explanatory variables. It is denoted as  $\eta = X\beta$ , where  $X$  represents the explanatory variables, and  $\beta$  represents the coefficients.
- **Link Function:** The link function,  $g$ , provides the relationship between the linear predictor and the mean of the distribution function. For a GLM, the mean of  $Y$  is related to the linear predictor through the link function as  $\mu = g^{-1}(\eta)$ .

GLMs adapt to various data types through the selection of appropriate link functions and probability distributions. Here, we outline four special cases of GLM: normal regression, logistic regression, Poisson regression, and gamma regression.

- **Normal Regression (Linear Regression).** In normal regression, the response variable has a normal distribution. The identity link function ( $g(\mu) = \mu$ ) is typically used, making this case equivalent to classical linear regression.
  - **Use Case:** Modeling continuous data where residuals are normally distributed.
  - **Link Function:** Identity ( $g(\mu) = \mu$ )
  - **Distribution:** Normal
- **Logistic Regression.** Logistic regression is used for binary response variables. It employs the logit link function to model the probability that an observation falls into one of two categories.

## 8 Statistical Tests and Models

- **Use Case:** Binary outcomes (e.g., success/failure).
  - **Link Function:** Logit ( $g(\mu) = \log \frac{\mu}{1-\mu}$ )
  - **Distribution:** Binomial
- Poisson Regression. Poisson regression models count data using the Poisson distribution. It's ideal for modeling the rate at which events occur.
  - **Use Case:** Count data, such as the number of occurrences of an event.
  - **Link Function:** Log ( $g(\mu) = \log(\mu)$ )
  - **Distribution:** Poisson
- Gamma Regression. Gamma regression is suited for modeling positive continuous variables, especially when data are skewed and variance increases with the mean.
  - **Use Case:** Positive continuous outcomes with non-constant variance.
  - **Link Function:** Inverse ( $g(\mu) = \frac{1}{\mu}$ )
  - **Distribution:** Gamma

Each GLM variant addresses specific types of data and research questions, enabling precise modeling and inference based on the underlying data distribution.

A logistic regression can be fit with `statsmodels.api.glm`.

To demonstrate the validation of logistic regression models, we first create a simulated dataset with binary outcomes. This setup involves generating logistic probabilities and then drawing binary outcomes based on these probabilities.

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
```

## 8.2 Statistical Modeling

```
# Set seed for reproducibility
np.random.seed(42)

# Create a DataFrame with random features named `simdat`
simdat = pd.DataFrame(np.random.randn(1000, 5), columns=['x1', 'x2', 'x3', 'x4', 'x5'])

# Calculating the linear combination of inputs plus an intercept
eta = simdat.dot([2, 2, 2, 2, 2]) - 5

# Applying the logistic function to get probabilities using statsmodels' logit link
p = sm.families.links.Logit().inverse(eta)

# Generating binary outcomes based on these probabilities and adding them to `simdat`
simdat['yb'] = np.random.binomial(1, p, p.size)

# Display the first few rows of the dataframe
print(simdat.head())
```

	x1	x2	x3	x4	x5	yb
0	0.496714	-0.138264	0.647689	1.523030	-0.234153	0
1	-0.234137	1.579213	0.767435	-0.469474	0.542560	0
2	-0.463418	-0.465730	0.241962	-1.913280	-1.724918	0
3	-0.562288	-1.012831	0.314247	-0.908024	-1.412304	0
4	1.465649	-0.225776	0.067528	-1.424748	-0.544383	0

Fit a logistic regression for y1b with the formula interface.

```
import statsmodels.formula.api as smf

# Specify the model formula
formula = 'yb ~ x1 + x2 + x3 + x4 + x5'
```

```
# Fit the logistic regression model using glm and a formula
fit = smf.glm(formula=formula, data=simdat, family=sm.families.Binomial()).f

# Print the summary of the model
print(fit.summary())
```

Generalized Linear Model Regression Results						
=====						
Dep. Variable:	yb	No. Observations:				100
Model:	GLM	Df Residuals:				99
Model Family:	Binomial	Df Model:				
Link Function:	Logit	Scale:				1.000
Method:	IRLS	Log-Likelihood:				-159.7
Date:	Mon, 07 Oct 2024	Deviance:				318.7
Time:	13:17:05	Pearson chi2:				1.47e+
No. Iterations:	8	Pseudo R-squ. (CS):				0.41
Covariance Type:	nonrobust					
=====						
	coef	std err	z	P> z	[0.025	0.975]
-----						
Intercept	-5.0186	0.392	-12.796	0.000	-5.787	-4.239
x1	1.9990	0.211	9.471	0.000	1.585	2.413
x2	2.1058	0.214	9.853	0.000	1.687	2.524
x3	1.9421	0.210	9.260	0.000	1.531	2.353
x4	2.1504	0.232	9.260	0.000	1.695	2.605
x5	2.0603	0.221	9.313	0.000	1.627	2.493
=====						

8.3 Validating the Results of Logistic Regression

Validating the performance of logistic regression models is crucial to assess their effectiveness and reliability. This section explores key metrics used

### 8.3 Validating the Results of Logistic Regression

to evaluate the performance of logistic regression models, starting with the confusion matrix, then moving on to accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC). Using simulated data, we will demonstrate how to calculate and interpret these metrics using Python.

#### 8.3.1 Confusion Matrix

The confusion matrix is a fundamental tool used for calculating several other classification metrics. It is a table used to describe the performance of a classification model on a set of data for which the true values are known. The matrix displays the actual values against the predicted values, providing insight into the number of correct and incorrect predictions.

Actual	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

[https://en.wikipedia.org/wiki/Confusion\\_matrix](https://en.wikipedia.org/wiki/Confusion_matrix)

Four entries in the confusion matrix:

- True Positive (TP): The cases in which the model correctly predicted the positive class.
- False Positive (FP): The cases in which the model incorrectly predicted the positive class (i.e., the model predicted positive, but the actual class was negative).
- True Negative (TN): The cases in which the model correctly predicted the negative class.
- False Negative (FN): The cases in which the model incorrectly predicted the negative class (i.e., the model predicted negative, but the actual class was positive).

Four rates from the confusion matrix with actual (row) margins:

## 8 Statistical Tests and Models

- True positive rate (TPR):  $TP / (TP + FN)$ . Also known as sensitivity.
- False negative rate (FNR):  $FN / (TP + FN)$ . Also known as miss rate.
- False positive rate (FPR):  $FP / (FP + TN)$ . Also known as false alarm, fall-out.
- True negative rate (TNR):  $TN / (FP + TN)$ . Also known as specificity.

Note that TPR and FPR do not add up to one. Neither do FNR and FPR.

- Positive predictive value (PPV):  $TP / (TP + FP)$ . Also known as precision.
- False discovery rate (FDR):  $FP / (TP + FP)$ .
- False omission rate (FOR):  $FN / (FN + TN)$ .
- Negative predictive value (NPV):  $TN / (FN + TN)$ .

Note that PPV and NP do not add up to one.

### 8.3.2 Accuracy

Accuracy measures the overall correctness of the model and is defined as the ratio of correct predictions (both positive and negative) to the total number of cases examined.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

- Imbalanced Classes: Accuracy can be misleading if there is a significant imbalance between the classes. For instance, in a dataset where 95% of the samples are of one class, a model that naively predicts the majority class for all instances will still achieve 95% accuracy, which does not reflect true predictive performance.



### 8.3 Validating the Results of Logistic Regression

- **Misleading Interpretations:** High overall accuracy might hide the fact that the model is performing poorly on a smaller, yet important, segment of the data.

#### 8.3.3 Precision

Precision (or PPV) measures the accuracy of positive predictions. It quantifies the number of correct positive predictions made.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

- **Neglect of False Negatives:** Precision focuses solely on the positive class predictions. It does not take into account false negatives (instances where the actual class is positive but predicted as negative). This can be problematic in cases like disease screening where missing a positive case (disease present) could be dangerous.
- **Not a Standalone Metric:** High precision alone does not indicate good model performance, especially if recall is low. This situation could mean the model is too conservative in predicting positives, thus missing out on a significant number of true positive instances.

#### 8.3.4 Recall

Recall (Sensitivity or TPR) measures the ability of a model to find all relevant cases (all actual positives).

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

- **Neglect of False Positives:** Recall does not consider false positives (instances where the actual class is negative but predicted as positive). High recall can be achieved at the expense of precision, leading to a large number of false positives which can be costly or undesirable in certain contexts, such as in spam detection.

- Trade-off with Precision: Often, increasing recall decreases precision. This trade-off needs to be managed carefully, especially in contexts where both false positives and false negatives carry significant costs or risks.

### 8.3.5 F-beta Score

The F-beta score is a weighted harmonic mean of precision and recall, taking into account a  $\beta$  parameter such that recall is considered  $\beta$  times as important as precision:

$$(1 + \beta^2) \frac{\text{precision} \cdot \text{recall}}{\beta^2 \text{precision} + \text{recall}}.$$

See [stackexchange post](#) for the motivation of  $\beta^2$  instead of just  $\beta$ .

The F-beta score reaches its best value at 1 (perfect precision and recall) and worst at 0.

If reducing false negatives is more important (as might be the case in medical diagnostics where missing a positive diagnosis could be critical), you might choose a beta value greater than 1. If reducing false positives is more important (as in spam detection, where incorrectly classifying an email as spam could be inconvenient), a beta value less than 1 might be appropriate.

The F1 Score is a specific case of the F-beta score where beta is 1, giving equal weight to precision and recall. It is the harmonic mean of Precision and Recall and is a useful measure when you seek a balance between Precision and Recall and there is an uneven class distribution (large number of actual negatives).

### 8.3.6 Receiver Operating Characteristic (ROC) Curve

The **Receiver Operating Characteristic (ROC) curve** is a plot that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. It shows the trade-off between the TPR and FPR. The ROC plots TPR against FPR as the decision threshold is varied. It can be particularly useful in evaluating the performance of classifiers when the class distribution is imbalanced,

The **Area Under the ROC Curve (AUC)** is a scalar value that summarizes the performance of a classifier. It measures the total area underneath the ROC curve, providing a single metric to compare models. The value of AUC ranges from 0 to 1:

- **AUC = 1:** A perfect classifier, which perfectly separates positive and negative classes.
- **AUC = 0.5:** A classifier that performs no better than random chance.
- **AUC < 0.5:** A classifier performing worse than random.

The AUC value provides insight into the model's ability to discriminate between positive and negative classes across all possible threshold values.

### 8.3.7 Demonstration

Let's apply these metrics to the `simdat` dataset to understand their practical implications. We will fit a logistic regression model, make predictions, and then compute accuracy, precision, and recall.

```
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import (
    accuracy_score, precision_score, recall_score, confusion_matrix,
```

## 8 Statistical Tests and Models

```
f1_score, roc_curve, auc
)
import matplotlib.pyplot as plt
from sklearn.datasets import make_classification

# Generate synthetic data
X, y = make_classification(n_samples=1000, n_features=20, random_state=42)

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)

# Fit the logistic regression model
model = LogisticRegression()
model.fit(X_train, y_train)

# Predict labels on the test set
y_pred = model.predict(X_test)

# Get predicted probabilities for ROC curve and AUC
y_scores = model.predict_proba(X_test)[:, 1] # Probability for the positive class

# Compute confusion matrix
cm = confusion_matrix(y_test, y_pred)

# Calculate accuracy, precision, and recall
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)

# Print confusion matrix and metrics
print("Confusion Matrix:\n", cm)
print(f"Accuracy: {accuracy:.2f}")
print(f"Precision: {precision:.2f}")
```

### 8.3 Validating the Results of Logistic Regression

```
print(f"Recall: {recall:.2f}")
```

Confusion Matrix:

```
[[104  11]
```

```
 [ 26 109]]
```

Accuracy: 0.85

Precision: 0.91

Recall: 0.81

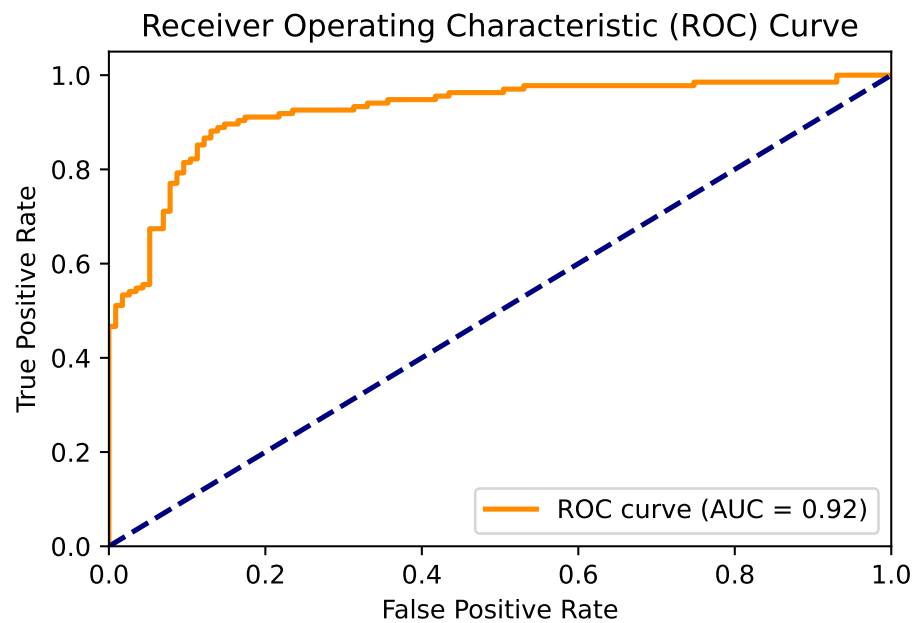
By varying threshold, one can plot the whole ROC curve.

```
# Compute ROC curve and AUC
fpr, tpr, thresholds = roc_curve(y_test, y_scores)
roc_auc = auc(fpr, tpr)

# Print AUC
print(f"AUC: {roc_auc:.2f}")

# Plot ROC curve
plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--') # Diagonal line (random classi
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc="lower right")
plt.show()
```

AUC: 0.92



We could pick the best threshold that optimizes F1-score/

```
# Compute F1 score for each threshold
f1_scores = []
for thresh in thresholds:
    y_pred_thresh = (y_scores >= thresh).astype(int) # Apply threshold to g
    f1 = f1_score(y_test, y_pred_thresh)
    f1_scores.append(f1)

# Find the best threshold (the one that maximizes F1 score)
best_thresh = thresholds[np.argmax(f1_scores)]
best_f1 = max(f1_scores)

# Print the best threshold and corresponding F1 score
print(f"Best threshold: {best_thresh:.4f}")
```

```
print(f"Best F1 score: {best_f1:.2f}")
```

Best threshold: 0.3960

Best F1 score: 0.89

## 8.4 LASSO Logistic Models

The **Least Absolute Shrinkage and Selection Operator** (LASSO) (Tibshirani 1996), is a regression method that performs both **variable selection** and **regularization**. LASSO imposes an L1 penalty on the regression coefficients, which has the effect of shrinking some coefficients exactly to zero. This results in simpler, more interpretable models, especially in situations where the number of predictors exceeds the number of observations.

### 8.4.1 Theoretical Formulation of the Problem

The objective function for LASSO logistic regression can be expressed as,

$$\min_{\beta} \left\{ -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)] + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

where:

- $\hat{p}_i = \frac{1}{1+e^{-X_i\beta}}$  is the predicted probability for the  $i$ -th sample.
- $y_i$  represents the actual class label (binary: 0 or 1).
- $X_i$  is the feature vector for the  $i$ -th observation.
- $\beta$  is the vector of model coefficients (including the intercept).

## 8 Statistical Tests and Models

- $\lambda$  is the regularization parameter that controls the trade-off between model fit and sparsity (higher  $\lambda$ ) encourages sparsity by shrinking more coefficients to zero).

The lasso penalty encourages the sum of the absolute values of the coefficients to be small, effectively shrinking some coefficients to zero. This results in sparser solutions, simplifying the model and reducing variance without substantial increase in bias.

Practical benefits of LASSO:

- **Dimensionality Reduction:** LASSO is particularly useful when the number of features ( $p$ ) is large, potentially even larger than the number of observations ( $n$ ), as it automatically reduces the number of features.
- **Preventing Overfitting:** The L1 penalty helps prevent overfitting by constraining the model, especially when ( $p$ ) is large or there is multicollinearity among features.
- **Interpretability:** By selecting only the most important features, LASSO makes the resulting model more interpretable, which is valuable in fields like bioinformatics, economics, and social sciences.

### 8.4.2 Solution Path

To illustrate the effect of the lasso penalty in logistic regression, we can plot the solution path of the coefficients as a function of the regularization parameter ( $\lambda$ ). This demonstration will use a simulated dataset to show how increasing ( $\lambda$ ) leads to more coefficients being set to zero.

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import make_classification
```



## 8.4 LASSO Logistic Models

```
# Function to create a lambda grid
def compute_lambda_max_logistic(X, y):
    """
    Compute lambda_max for logistic regression (smallest lambda at which all coefficients are zero)

    Parameters:
    - X: Feature matrix (numpy array or similar)
    - y: Target vector (numpy array or similar)

    Returns:
    - lambda_max: The largest lambda for which all coefficients are zero
    """
    # Standardize the features
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)

    # Initial prediction probability (0.5 since all coefficients are zero)
    p = 0.5 * np.ones_like(y)

    # Calculate gradient at zero coefficients
    gradient = np.dot(X_scaled.T, p - y)

    # Lambda max is the maximum of the absolute gradients divided by the number of samples
    lambda_max = np.max(np.abs(gradient)) / X_scaled.shape[0]

    return lambda_max

def create_lambda_grid(X, y, lambda_min_ratio, num_lambdas=100):
    """
    Create a logarithmically spaced lambda grid for L1-penalized logistic regression.

    Parameters:
```

```

- X: Feature matrix (numpy array or similar)
- y: Target vector (numpy array or similar)
- lambda_min_ratio: The ratio of the smallest lambda in the grid to lambda_max
- num_lambdas: Number of lambda values to generate in the grid

Returns:
- lambdas: Array of lambda values from lambda_max to lambda_min, logarithmic scale
- Cs: Array of C values for use in scikit-learn, where C = 1/lambda
"""
# Compute lambda_max for logistic regression
lambda_max = compute_lambda_max_logistic(X, y)

# Compute lambda_min using the lambda_min_ratio
lambda_min = lambda_max * lambda_min_ratio

# Generate a sequence of lambdas from lambda_max to lambda_min, logarithmic scale
lambdas = np.logspace(np.log10(lambda_max), np.log10(lambda_min), num=num_lambdas)

# Convert lambdas to C values for use in scikit-learn (C = 1/lambda)
Cs = 1 / lambdas

return lambdas, Cs

# Generate synthetic data
X, y = make_classification(n_samples=100, n_features=20, n_informative=2, random_state=1)

# Generate lambda and C grid
lambda_min_ratio = 0.01
lambdas, Cs = create_lambda_grid(X, y, lambda_min_ratio)

# Fit logistic regression model for each C and store coefficients
coefficients = []
scaler = StandardScaler()

```

## 8.4 LASSO Logistic Models

```
X_scaled = scaler.fit_transform(X)

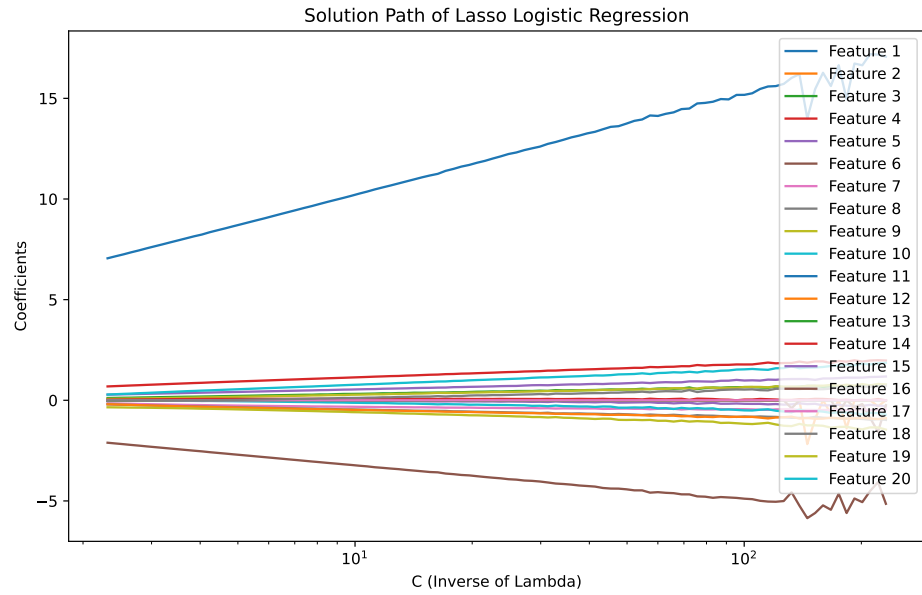
for C in Cs:
    model = LogisticRegression(penalty='l1', C=C, solver='liblinear', max_iter=10000)
    model.fit(X_scaled, y)
    coefficients.append(model.coef_[0])

# Plot solution path against C (inverse of lambda)
coefficients = np.array(coefficients)
plt.figure(figsize=(10, 6))

for i in range(coefficients.shape[1]):
    plt.plot(Cs, coefficients[:, i], label=f'Feature {i+1}')

plt.xscale('log')
plt.xlabel('C (Inverse of Lambda)')
plt.ylabel('Coefficients')
plt.title('Solution Path of Lasso Logistic Regression')
plt.legend()
plt.show()
```

## 8 Statistical Tests and Models



### 8.4.3 Selection the Tuning Parameter

In logistic regression with LASSO regularization, selecting the optimal value of the regularization parameter  $C$  (the inverse of  $\lambda$ ) is crucial to balancing the model's bias and variance. A small  $C$  value (large  $\lambda$ ) increases the regularization effect, shrinking more coefficients to zero and simplifying the model. Conversely, a large  $C$  (small  $\lambda$ ) allows the model to fit the data more closely.

The best way to select the optimal  $C$  is through **cross-validation**. In cross-validation, the dataset is split into several folds, and the model is trained on some folds while evaluated on the remaining fold. This process is repeated for each fold, and the results are averaged to ensure the model generalizes well to unseen data. The  $C$  value that results in the best performance is selected.

## 8.4 LASSO Logistic Models

The performance metric used in cross-validation can vary based on the task. Common metrics include:

- **Log-loss:** Measures how well the predicted probabilities match the actual outcomes.
- **Accuracy:** Measures the proportion of correctly classified instances.
- **F1-Score:** Balances precision and recall, especially useful for imbalanced classes.
- **AUC-ROC:** Evaluates how well the model discriminates between the positive and negative classes.

In Python, the `LogisticRegressionCV` class from `scikit-learn` automates cross-validation for logistic regression. It evaluates the model's performance for a range of  $C$  values and selects the best one.

```
import numpy as np
from sklearn.linear_model import LogisticRegressionCV
from sklearn.model_selection import train_test_split
from sklearn.datasets import make_classification
from sklearn.metrics import accuracy_score

# Generate synthetic data
X, y = make_classification(n_samples=1000, n_features=20, random_state=42)

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)

# Initialize LogisticRegressionCV with L1 penalty for Lasso and cross-validation
log_reg_cv = LogisticRegressionCV(
    Cs=np.logspace(-4, 4, 20), # Range of C values (inverse of lambda)
    cv=5, # 5-fold cross-validation
    penalty='l1', # Lasso regularization (L1 penalty)
    solver='liblinear', # Solver for L1 regularization
    scoring='accuracy', # Optimize for accuracy
```

## 8 Statistical Tests and Models

```
        max_iter=10000                # Ensure convergence
    )

    # Train the model with cross-validation
    log_reg_cv.fit(X_train, y_train)

    # Best C value (inverse of lambda)
    print(f"Best C value: {log_reg_cv.C_[0]}")

    # Evaluate the model on the test set
    y_pred = log_reg_cv.predict(X_test)
    test_accuracy = accuracy_score(y_test, y_pred)
    print(f"Test Accuracy: {test_accuracy:.2f}")

    # Display the coefficients of the best model
    print("Model Coefficients:\n", log_reg_cv.coef_)
```

Best C value: 0.08858667904100823

Test Accuracy: 0.86

Model Coefficients:

[[ 0.	0.	0.05552448	0.	0.	1.90889734
0.	0.	0.	0.	0.0096863	0.23541942
0.	0.	-0.0268928	0.	0.	0.
0.	0.	]]			

## 9 Supervised Learning

### 9.1 Introduction

Supervised learning uses labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross validation process.

In contrast, unsupervised learning uses unlabeled data to discover patterns that help solve for clustering or association problems. This is particularly useful when subject matter experts are unsure of common properties within a data set.

### 9.2 Classification vs Regression

- Classification: outcome variable is categorical
- Regression: outcome variable is continuous
- Both problems can have many covariates (predictors/features)

#### 9.2.1 Regression metrics

- Mean squared error (MSE)
- Mean absolute error (MAE)

## 9.2.2 Classification metrics

### 9.2.2.1 Confusion matrix

[https://en.wikipedia.org/wiki/Confusion\\_matrix](https://en.wikipedia.org/wiki/Confusion_matrix)

Four entries in the confusion matrix:

- TP: number of true positives
- FN: number of false negatives
- FP: number of false positives
- TN: number of true negatives

Four rates from the confusion matrix with actual (row) margins:

- TPR:  $TP / (TP + FN)$ . Also known as sensitivity.
- FNR:  $TN / (TP + FN)$ . Also known as miss rate.
- FPR:  $FP / (FP + TN)$ . Also known as false alarm, fall-out.
- TNR:  $TN / (FP + TN)$ . Also known as specificity.

Note that TPR and FPR do not add up to one. Neither do FNR and FPR.

Four rates from the confusion matrix with predicted (column) margins:

- PPV:  $TP / (TP + FP)$ . Also known as precision.
- FDR:  $FP / (TP + FP)$ .
- FOR:  $FN / (FN + TN)$ .
- NPV:  $TN / (FN + TN)$ .

### 9.2.2.2 Measure of classification performance

Measures for a given confusion matrix:

- Accuracy:  $(TP + TN) / (P + N)$ . The proportion of all corrected predictions. Not good for highly imbalanced data.



## 9.2 Classification vs Regression

- Recall (sensitivity/TPR):  $TP / (TP + FN)$ . Intuitively, the ability of the classifier to find all the positive samples.
- Precision:  $TP / (TP + FP)$ . Intuitively, the ability of the classifier not to label as positive a sample that is negative.
- F-beta score: Harmonic mean of precision and recall with  $\beta$  chosen such that recall is considered  $\beta$  times as important as precision,

$$(1 + \beta^2) \frac{\text{precision} \cdot \text{recall}}{\beta^2 \text{precision} + \text{recall}}$$

See stackexchange post for the motivation of  $\beta^2$ .

When classification is obtained by dichotomizing a continuous score, the receiver operating characteristic (ROC) curve gives a graphical summary of the FPR and TPR for all thresholds. The ROC curve plots the TPR against the FPR at all thresholds.

- Increasing from  $(0, 0)$  to  $(1, 1)$ .
- Best classification passes  $(0, 1)$ .
- Classification by random guess gives the 45-degree line.
- Area between the ROC and the 45-degree line is the Gini coefficient, a measure of inequality.
- Area under the curve (AUC) of ROC thus provides an important metric of classification results.

### 9.2.3 Cross-validation

- Goal: strike a bias-variance tradeoff.
- K-fold: hold out each fold as testing data.
- Scores: minimized to train a model

Cross-validation is an important measure to prevent over-fitting. Good in-sample performance does not necessarily mean good out-sample performance. A general work flow in model selection with cross-validation is as follows.

## 9 Supervised Learning

- Split the data into training and testing
- For each candidate model  $m$  (with possibly multiple tuning parameters)
  - Fit the model to the training data
  - Obtain the performance measure  $f(m)$  on the testing data (e.g., CV score, MSE, loss, etc.)
- Choose the model  $m^* = \arg \max_m f(m)$ .

## 10 Exercises

1. **Quarto and Git setup** Quarto and Git are two important tools for data science. Get familiar with them through the following tasks. Please use the `templates/hw.qmd` template.
  1. Install Quarto onto your computer following the instructions of Get Started. Document the obstacles you encountered and how you overcame them.
  2. Pick a tool of your choice (e.g., VS Code, Jupyter Notebook, Emacs, etc.), follow the instructions to reproduce the example of line plot on polar axis.
  3. Render the homework into a pdf file and put the file into a release in your GitHub repo. Document any obstacles you have and how you overcome them.
2. **Git basics and GitHub setup** Learn the Git basics and set up an account on GitHub if you do not already have one. Practice the tips on Git in the notes. By going through the following tasks, ensure your repo has at least 10 commits, each with an informative message. Regularly check the status of your repo using `git status`. The specific tasks are:
  1. Clone the class notes repo to an appropriate folder on your computer.
  2. Add all the files to your designated homework repo from GitHub Classroom and work on that repo for the rest of the problem.
  3. Add your name and wishes to the Wishlist; commit.
  4. Remove the **Last, First** entry from the list; commit.

## 10 Exercises

5. Create a new file called `add.qmd` containing a few lines of texts; commit.
  6. Remove `add.qmd` (pretending that this is by accident); commit.
  7. Recover the accidentally removed file `add.qmd`; add a long line (a paragraph without a hard break); add a short line (under 80 characters); commit.
  8. Change one word in the long line and one word in the short line; use `git diff` to see the difference from the last commit; commit.
  9. Play with other git operations and commit.
3. **Contributing to the Class Notes** To contribute to the classnotes, you need to have a working copy of the sources on your computer. Document the following steps in a `qmd` file as if you are explaining them to someone who want to contribute too.
1. Create a fork of the notes repo into your own GitHub account.
  2. Clone it to an appropriate folder on your computer.
  3. Render the classnotes on your computer; document the obstacles and solutions.
  4. Make a new branch (and name it appropriately) to experiment with your changes.
  5. Checkout your branch and add your wishes to the wish list; commit with an informative message; and push the changes to your GitHub account.
  6. Make a pull request to class notes repo from your fork at GitHub. Make sure you have clear messages to document the changes.
4. **Monty Hall** Write a function to demonstrate the Monty Hall problem through simulation. The function takes two arguments `nddoors` and `ntrials`, representing the number of doors in the experiment and the number of trials in a simulation, respectively. The function should return the proportion of wins for both the switch and

no-switch strategy. Apply your function with 3 doors and 5 doors, both with 1000 trials. Include sufficient text around the code to explain your them.

5. **Approximating  $\pi$**  Write a function to do a Monte Carlo approximation of  $\pi$ . The function takes a Monte Carlo sample size  $n$  as input, and returns a point estimate of  $\pi$  and a 95% confidence interval. Apply your function with sample size 1000, 2000, 4000, and 8000. Repeat the experiment 1000 times for each sample size and check the empirical probability that the confidence intervals cover the true value of  $\pi$ . Comment on the results.
6. **Google Billboard Ad** Find the first 10-digit prime number occurring in consecutive digits of  $e$ . This was a Google recruiting ad.
7. **Game 24** The math game 24 is one of the addictive games among number lovers. With four randomly selected cards from a deck of poker cards, use all four values and elementary arithmetic operations ( $+-\times/$ ) to come up with 24. Let  $\square$  be one of the four numbers. Let  $\circ$  represent one of the four operators. For example,

$$(\square \circ \square) \circ (\square \circ \square)$$

is one way to group the the operations.

1. List all the possible ways to group the four numbers.
  2. How many possibly ways are there to check for a solution?
  3. Write a function to solve the problem in a brutal force way. The inputs of the function are four numbers. The function returns a list of solutions. Some of the solutions will be equivalent, but let us not worry about that for now.
8. **NYC Crash Data Cleaning** The NYC motor vehicle collisions data with documentation is available from NYC Open Data. The raw data needs some cleaning.

## 10 Exercises

1. Use the filter from the website to download the crash data of the week of June 30, 2024 in CSV format; save it under a directory `data` with an informative name (e.g., `nyccrashes_2024w0630_by20240916.csv`); read the data into a Panda data frame with careful handling of the date time variables.
2. Clean up the variable names. Use lower cases and replace spaces with underscores.
3. Get the basic summaries of each variables: missing percentage; descriptive statistics for continuous variables; frequency tables for discrete variables.
4. Are there invalid `longitude` and `latitude` in the data? If so, replace them with NA.
5. Are there `zip_code` values that are not legit NYC zip codes? If so, replace them with NA.
6. Are there missing in `zip_code` and `borough`? Do they always co-occur?
7. Are there cases where `zip_code` and `borough` are missing but the geo codes are not missing? If so, fill in `zip_code` and `borough` using the geo codes.
8. Is it redundant to keep both `location` and the `longitude/latitude` at the NYC Open Data server?
9. Check the frequency of `crash_time` by hour. Is there a matter of bad luck at exactly midnight? How would you interpret this?
10. Are the number of persons killed/injured the summation of the numbers of pedestrians, cyclist, and motorists killed/injured? If so, is it redundant to keep these two columns at the NYC Open Data server?
11. Print the whole frequency table of `contributing_factor_vehicle_1`. Convert lower cases to uppercases and check the frequencies again.
12. Provided an opportunity to meet the data provider, what suggestions would you make based on your data exploration experience?

9. **NYC Crash Data Exploration** Except for the first question, use the cleaned crash data in feather format.

1. Construct a contingency table for missing in geocode (latitude and longitude) by borough. Is the missing pattern the same across boroughs? Formulate a hypothesis and test it.
2. Construct a **hour** variable with integer values from 0 to 23. Plot the histogram of the number of crashes by **hour**. Plot it by borough.
3. Overlay the locations of the crashes on a map of NYC. The map could be a static map or Google map.
4. Create a new variable **severe** which is one if the number of persons injured or deaths is 1 or more; and zero otherwise. Construct a cross table for **severe** versus borough. Is the severity of the crashes the same across boroughs? Test the null hypothesis that the two variables are not associated with an appropriate test.
5. Merge the crash data with the zip code database.
6. Fit a logistic model with **severe** as the outcome variable and covariates that are available in the data or can be engineered from the data. For example, zip code level covariates can be obtained by merging with the zip code database; crash hour; number of vehicles involved.

10. **NYC Crash severity modeling** Using the cleaned NYC crash data, merged with zipcode level information, predict **severe** of a crash.

1. Set random seed to 1234. Randomly select 20% of the crashes as testing data and leave the rest 80% as training data.
2. Fit a logistic model on the training data and validate the performance on the testing data. Explain the confusion matrix result from the testing data. Compute the F1 score.

## 10 Exercises

3. Fit a logistic model on the training data with  $L_1$  regularization. Select the tuning parameter with 5-fold cross-validation in F1 score
  4. Apply the regularized logistic regression to predict the severity of the crashes in the testing data. Compare the performance of the two logistic models in terms of accuracy, precision, recall, F1-score, and AUC.
11. The NYC Open Data of 311 Service Requests contains all requests from 2010 to present. We consider a subset of it with request time between 00:00:00 01/15/2023 and 24:00:00 01/21/2023. The subset is available in CSV format as `data/nyc311_011523-012123_by022023.csv`. Read the data dictionary to understand the meaning of the variables,
1. Clean the data: fill missing fields as much as possible; check for obvious data entry errors (e.g., can `Closed Date` be earlier than `Created Date`?); summarize your suggestions to the data curator in several bullet points.
  2. Remove requests that are not made to NYPD and create a new variable `duration`, which represents the time period from the `Created Date` to `Closed Date`. Note that `duration` may be censored for some requests. Visualize the distribution of uncensored `duration` by weekdays/weekend and by borough, and test whether the distributions are the same across weekdays/weekends of their creation and across boroughs.
  3. Define a binary variable `over3h` which is 1 if `duration` is greater than 3 hours. Note that it can be obtained even for censored `duration`. Build a model to predict `over3h`. If your model has tuning parameters, justify their choices. Apply this model to the 311 requests of NYPD in the week of 01/22/2023. Assess the performance of your model.
  4. Now you know the data quite well. Come up with a research question of interest that can be answered by the data, which



could be analytics or visualizations. Perform the needed analyses and answer your question.

12. **Logistic Modeling** The response time to 311 service requests is a measure of civic service quality. Let us model the response time to 311 requests with complain type **Rodent**.

1. Compute the response time in hours. Note that some response will be missing because of unavailable closed date.
2. Compute a binary variable **over3d**, which is one if the response time is greater than 3 days, and zero otherwise. Note that this variable should have no missing values.
3. Use the package **uszipcode** to obtain the zip code level covariates such as median house income and median home value. Merge these variables to the rodent data.
4. Split the data at random into training (80%) and testing (20%). Build a logistic model to predict **over3d** on the training data, and validate the performance on the testing data.
5. Build a lasso logistic model to predict **over3d**, and justify your choice of the tuning parameter. Validate on the testing data.



## References

- Fayaz, Omer. 2022. “Principles of Visual Communication.”
- Grillo, H. M., and M. Enesi. 2022. “The Impact, Importance, Types, and Use of Non-Verbal Communication in Social Relations.” *Linguistics and Culture Review* 6: 291–307.
- Prabavathi, R., and P. C. Nagasubramani. 2018. “Effective Oral and Written Communication.” *Journal of Applied and Advanced Research* 10: 29–32.
- Pragmatic. 2024. “Communication Skills for Data Science.” <https://www.pragmaticinstitute.com%0A%20%20/resources/articles/data/communication-skills-for-data-science/>.
- Radovilsky, Z., V. Hegde, A. Acharya, and U. Uma. 2018. “Skills Requirements of Business Data Analytics and Data Science Jobs: A Comparative Analysis.” *Journal of Supply Chain and Operations Management* 16: 82–101.
- Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the LASSO.” *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88.
- Vandemeulebroecke, Marc, Mark Baillie, Alison Margolskee, and Baldur Magnusson. 2019. “Effective Visual Communication for the Quantitative Scientist.” *CPT Pharmacometrics Syst Pharmacol* 10: 705–19.
- VanderPlas, Jake. 2016. *Python Data Science Handbook: Essential Tools for Working with Data*. O’Reilly Media, Inc.
- Yer, Simona. 2018. “Verbal Communication as a Two-Way Process in Connecting People.” *Social Science Research Network*.

