Exploratory Data Analysis & Data Preparation

This assignment is a chance to explore your data. Be creative with your plots!

Contents

Exploring the Data

Answer (at least) the following questions:

- What is your outcome variable(s)? How well does it measure the outcome you are interested in? How does it relate to your expectations?

Our outcome variables are whether or not a resume gets passed to the next round. This directly relates to the outcome we are interested in, which is to determine whether hiring algorithms are biased. This will provide a concrete measure of our expectations.

- What are your key explanatory variables?

Technical skills, years of experience, professional experience, number of past jobs, gender, masculine words, feminine words

In addition, create a table of summary statistics for the variables you are planning to use in your analysis.

You can explore any other questions you see fit for your project.

Data Visualization

You must include at least 4 visualizations of your data made in Python. You must include your outcome variable in at least two plots and your key explanatory variable in at least two of these plots. You must use visualizations that are *appropriate* for the data type (categorical vs numeric, continuous vs discrete) of your outcome and explanatory variables. For example, you should not use a histogram to plot a categorical variable.

Data Preparation and Cleaning

Answer the following question:

- What data cleaning did you have to do?

To clean the data, we separated the resume data into more specific column categories. First, we separated the data by “summary”, “skill level” and "technical skills”. Then, we started extracting certain details such as number of past jobs and number of years of experience.

- How did you prepare the data?

We scraped 600 resumes from HireItPeople.com under the Java Developers/Architects

Resumes page. Hire IT People provides thousands of resumes within industries for employers to view. We created the resume.csv file by opening each link that holds a resume and then added the text to the csv. Once we had a csv containing the url link and the long resume text, we were able to begin separating the different variables into separate columns.

- Are you deciding to exclude any observations? If so, why?

We might remove some filler information, such as repetitive language and exclude resumes with the "professional summary" section (3 entries)

- Did you have to create any new variables from existing variables? If so, how and why?

We randomly assigned gender (male/female) to each entry. We also added columns for years of experience and number of previous jobs held, which were pulled from the resume descriptions. In addition, we added our own data for awards to relate to gender, specifically for females. We added awards that highlighted their role as a "woman in STEM".

- How Did We Pick Gender-Coded Words?

We used Gender Decoder which was inspired by a research paper written by Danielle Gaucher, Justin Friesen, and Aaron C. Kay back in 2011. In this paper, *Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality, "*researchers showed job adverts which included different kinds of gender-coded language to men and women and recorded how appealing the jobs seemed and how much the participants felt that they 'belonged' in that occupation."

- What STEM Awards did you use?

We used for female awards:

- Women Tech Award Recipient: *For women in computer science and related fields*
- Women of ENIAC Computer Pioneer Award Recipient: *recipients are individuals who have made significant contributions to the computer industry*
- Zonta Women in STEM Award Recipient: *Honors women between 18-35 for their contributions to STEM.*
- Women of Colour in STEM Award Recipient: *A new award that launched in 2024 to recognize the achievements of women of color in STEM.*
- Grace Murray Hopper Award: *For young computer professionals who made a significant technical or service contribution.*

For male awards:

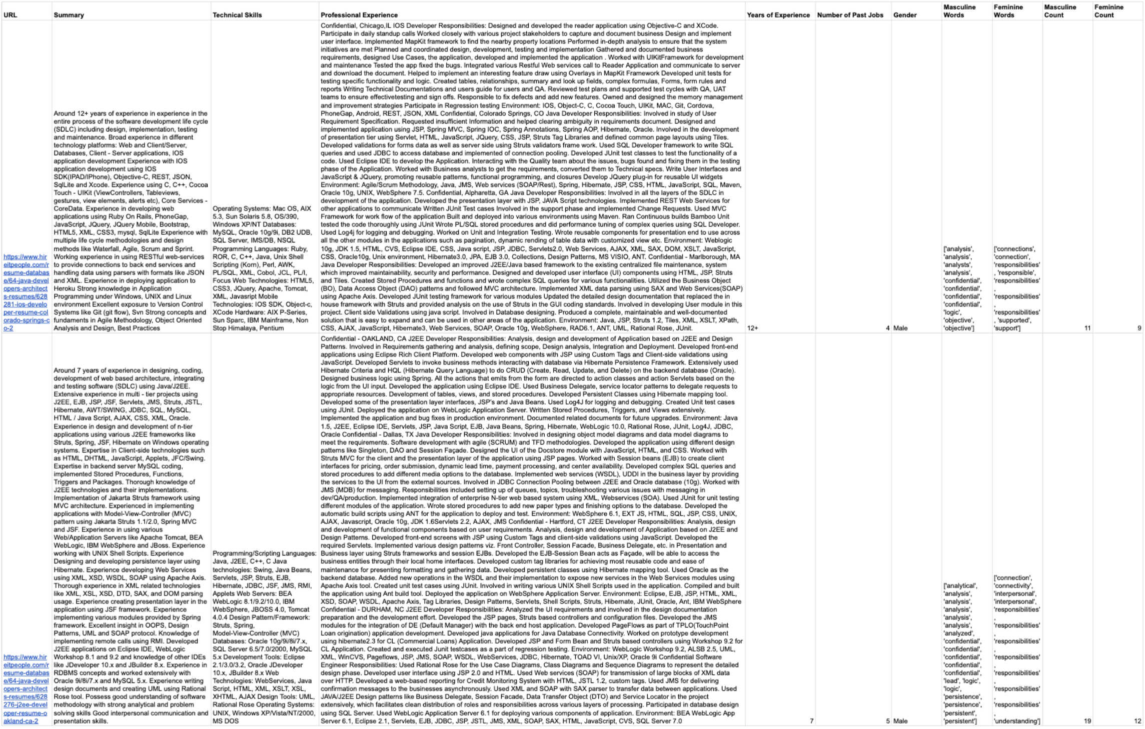
- ACM Prize in Computing: *Honors early-to-mid career innovators in computing.*
- ACM Software System Award: *For influential software systems with lasting impact.*
- The Java Community Process Award Recipient: *awards for excellence in Java standards development, such as "Member/Participant of the Year" and "Outstanding Spec Lead"*

Number of past jobs
Years of experience
Binary columns of skills
Add gender column based off of gender language decoder
https://gender-decoder.katmatfield.com/about
https://adminvc.ucla.edu/equity/hiring-guide/gender-decoder

Example Entry Data

Each data entry comes from a URL like this. It has a summary with past experiences, technical skills, and descriptions of past professional experience.

Below are two examples of a resume data point from our CSV file.



URL	Summary	Technical Skills	Professional Experience	Years of Experience	Number of Past Jobs	Gender	Masculine Words	Feminine Words	Masculine Count	Feminine Count
https://www.hireitpeople.com/resume-database/64-java-developers-architects-resumes/628281-ios-developer-resume-colorado-springs-co-2	Around 12+ years of experience in experience in the entire process of the software development life cycle (SDLC) including design, implementation, testing and maintenance. Broad experience in different technology platforms: Web and Client/Server, Databases, Client - Server applications, IOS application development Experience with IOS application development using IOS SDK(IPAD/IPhone), Objective-C, REST, JSON, SqlLite and Xcode. Experience using C, C++, Cocoa Touch - UIKit (ViewControllers, Tableviews, gestures, view elements, alerts etc), Core Services - CoreData. Experience in developing web applications using Ruby On Rails, PhoneGap, JavaScript, JQuery, JQuery Mobile, Bootstrap, HTML5, XML, CSS3, mysql, SqlLite Experience with multiple life cycle methodologies and design methods like Waterfall, Agile, Scrum and Sprint. Working experience in using RESTful web-services to provide connections to back end services and handling data using parsers with formats like JSON and XML. Experience in deploying application to Heroku Strong knowledge in Application Programming under Windows, UNIX and Linux environment Excellent exposure to Version Control Systems like Git (git flow), Svn Strong concepts and fundaments in Agile Methodology, Object Oriented Analysis and Design, Best Practices	Operating Systems: Mac OS, AIX 5.3, Sun Solaris 5.8, OS/390, Windows XP/NT Databases: MySQL, Oracle 10g/9i, DB2 UDB, SQL Server, IMS/DB, NSQL Programming Languages: Ruby, ROR, C, C++, Java, Unix Shell Scripting (Korn), Perl, AWK, PL/SQL, XML, Cobol, JCL, PL/I, Focus Web Technologies: HTML5, CSS3, JQuery, Apache, Tomcat, XML, Javasript Mobile Technologies: IOS SDK, Object-c, XCode Hardware: AIX P-Series, Sun Sparc, IBM Mainframe, Non Stop Himalaya, Pentium	Confidential, Chicago,IL IOS Developer Responsibilities: Designed and developed the reader application using Objective-C and XCode. Participate in daily standup calls Worked closely with various project stakeholders to capture and document business Design and implement user interface. Implemented MapKit framework to find the nearby property locations Performed in-depth analysis to ensure that the system initiatives are met Planned and coordinated design, development, testing and implementation Gathered and documented business requirements, designed Use Cases, the application, developed and implemented the application . Worked with UIKitFramework for development and maintenance Tested the app fixed the bugs. Integrated various Restful Web services call to Reader Application and communicate to server and download the document. Helped to implement an interesting feature draw using Overlays in MapKit Framework Developed unit tests for testing specific functionality and logic. Created tables, relationships, summary and look up fields, complex formulas, Forms, form rules and reports Writing Technical Documentations and users guide for users and QA. Reviewed test plans and supported test cycles with QA, UAT teams to ensure effectivetesting and sign offs. Responsible to fix defects and add new features. Owned and designed the memory management and improvement strategies Participate in Regression testing Environment: IOS, Object-C, C, Cocoa Touch, UIKit, MAC, Git, Cordova, PhoneGap, Android, REST, JSON, XML Confidential, Colorado Springs, CO Java Developer Responsibilities: Involved in study of User Requirement Specification. Requested insufficient Information and helped clearing ambiguity in requirements document. Designed and implemented application using JSP, Spring MVC, Spring IOC, Spring Annotations, Spring AOP, Hibernate, Oracle. Involved in the development of presentation tier using Servlet, HTML, JavaScript, JQuery, CSS, JSP, Struts Tag Libraries and defined common page layouts using Tiles. Developed validations for forms data as well as server side using Struts validators frame work. Used SQL Developer framework to write SQL queries and used JDBC to access database and implemented of connection pooling. Developed JUnit test classes to test the functionality of a code. Used Eclipse IDE to develop the Application. Interacting with the Quality team about the issues, bugs found and fixing them in the testing phase of the Application. Worked with Business analysts to get the requirements, converted them to Technical specs. Write User Interfaces and JavaScript & JQuery, promoting reusable patterns, functional programming, and closures Develop JQuery plug-in for reusable UI widgets Environment: Agile/Scrum Methodology, Java, JMS, Web services (SOAP/Rest), Spring, Hibernate, JSP, CSS, HTML, JavaScript, SQL, Maven, Oracle 10g, UNIX, WebSphere 7.5. Confidential, Alpharetta, GA Java Developer Responsibilities: Involved in all the layers of the SDLC in development of the application. Developed the presentation layer with JSP, JAVA Script technologies. Implemented REST Web Services for other applications to communicate Written JUnit Test cases Involved in the support phase and implemented Change Requests. Used MVC Framework for work flow of the application Built and deployed into various environments using Maven. Ran Continuous builds Bamboo Unit tested the code thoroughly using JUnit Wrote PL/SQL stored procedures and did performance tuning of complex queries using SQL Developer. Used Log4j for logging and debugging. Worked on Unit and Integration Testing. Wrote reusable components for presentation end to use across all the other modules in the applications such as pagination, dynamic rending of table data with customized view etc. Environment: Weblogic 10g, JDK 1.5, HTML, CVS, Eclipse IDE, CSS, Java script, JSP, JDBC, Servlets2.0, Web Services, AJAX, XML, SAX, DOM, XSLT, JavaScript, CSS, Oracle10g, Unix environment, Hibernate3.0, JPA, EJB 3.0, Collections, Design Patterns, MS VISIO, ANT. Confidential - Marlborough, MA Java Developer Responsibilities: Developed an improved J2EE/Java based framework to the existing centralized file maintenance, system which improved maintainability, security and performance. Designed and developed user interface (UI) components using HTML, JSP, Struts and Tiles. Created Stored Procedures and functions and wrote complex SQL queries for various functionalities. Utilized the Business Object (BO), Data Access Object (DAO) patterns and followed MVC architecture. Implemented XML data parsing using SAX and Web Services(SOAP) using Apache Axis. Developed JUnit testing framework for various modules Updated the detailed design documentation that replaced the in house framework with Struts and provided analysis on the use of Struts in the GUI coding standards. Involved in developing User module in this project. Client side Validations using java script. Involved in Database designing. Produced a complete, maintainable and well-documented solution that is easy to expand and can be used in other areas of the application. Environment: Java, JSP, Struts 1.2, Tiles, XML, XSLT, XPath, CSS, AJAX, JavaScript, Hibernate3, Web Services, SOAP, Oracle 10g, WebSphere, RAD6.1, ANT, UML, Rational Rose, JUnit.	12+	4	Male	['analysis', 'analysis', 'analysts', 'analysis', 'confidential', 'confidential', 'confidential', 'confidential', 'logic', 'objective', 'objective']	['connections', 'connection', 'responsibilities', 'responsible', 'responsibilities', 'responsibilities', 'responsibilities', 'supported', 'support']	11	9
https://www.hireitpeople.com/resume-database/64-java-developers-architects-resumes/628276-j2ee-developer-resume-oakland-ca-2	Around 7 years of experience in designing, coding, development of web based architecture, integrating and testing software (SDLC) using Java/J2EE. Extensive experience in multi - tier projects using J2EE, EJB, JSP, JSF, Servlets, JMS, Struts, JSTL, Hibernate, AWT/SWING, JDBC, SQL, MySQL, HTML / Java Script, AJAX, CSS, XML, Oracle. Experience in design and development of n-tier applications using various J2EE frameworks like Struts, Spring, JSF, Hibernate on Windows operating systems. Expertise in Client-side technologies such as HTML, DHTML, JavaScript, Applets, JFC/Swing. Expertise in backend server MySQL coding, implemented Stored Procedures, Functions, Triggers and Packages. Thorough knowledge of J2EE technologies and their implementations. Implementation of Jakarta Struts framework using MVC architecture. Experienced in implementing applications with Model-View-Controller (MVC) pattern using Jakarta Struts 1.1/2.0, Spring MVC and JSF. Experience in using various Web/Application Servers like Apache Tomcat, BEA WebLogic, IBM WebSphere and JBoss. Experience working with UNIX Shell Scripts. Experience Designing and developing persistence layer using Hibernate. Experience developing Web Services using XML, XSD, WSDL, SOAP using Apache Axis. Thorough experience in XML related technologies like XML, XSL, XSD, DTD, SAX, and DOM parsing usage. Experience creating presentation layer in the application using JSF framework. Experience implementing various modules provided by Spring framework. Excellent insight in OOPS, Design Patterns, UML and SOAP protocol. Knowledge of implementing remote calls using RMI. Developed J2EE applications on Eclipse IDE, WebLogic Workshop 8.1 and 9.2 and knowledge of other IDEs like JDeveloper 10.x and JBuilder 8.x. Experience in RDBMS concepts and worked extensively with Oracle 9i/8i/7.x and MySQL 5.x. Experience writing design documents and creating UML using Rational Rose tool. Possess good understanding of software methodology with strong analytical and problem solving skills Good interpersonal communication and presentation skills.	Programming/Scripting Languages: Java, J2EE, C++, C Java technologies: Swing, Java Beans, Servlets, JSP, Struts, EJB, Hibernate, JDBC, JSF, JMS, RMI, Applets Web Servers: BEA WebLogic 8.1/9.2/10.0, IBM WebSphere, JBOSS 4.0, Tomcat 4.0.4 Design Pattern/Framework: Struts, Spring, Model-View-Controller (MVC) Databases: Oracle 10g/9i/8i/7.x, SQL Server 6.5/7.0/2000, MySQL 5.x Development Tools: Eclipse 2.1/3.0/3.2, Oracle JDeveloper 10.x, JBuilder 8.x Web Technologies: WebServices, Java Script, HTML, XML, XSLT, XSL, XHTML, AJAX Design Tools: UML, Rational Rose Operating Systems: UNIX, Windows XP/Vista/NT/2000, MS DOS	Confidential - OAKLAND, CA J2EE Developer Responsibilities: Analysis, design and development of Application based on J2EE and Design Patterns. Involved in Requirements gathering and analysis, defining scope, Design analysis, Integration and Deployment. Developed front-end applications using Eclipse Rich Client Platform. Developed web components with JSP using Custom Tags and Client-side validations using JavaScript. Developed Servlets to invoke business methods interacting with database via Hibernate Persistence Framework. Extensively used Hibernate Criteria and HQL (Hibernate Query Language) to do CRUD (Create, Read, Update, and Delete) on the backend database (Oracle). Designed business logic using Spring. All the actions that emits from the form are directed to action classes and action Servlets based on the logic from the UI input. Developed the application using Eclipse IDE. Used Business Delegate, service locator patterns to delegate requests to appropriate resources. Development of tables, views, and stored procedures. Developed Persistent Classes using Hibernate mapping tool. Developed some of the presentation layer interfaces, JSP's and Java Beans. Used Log4J for logging and debugging. Created Unit test cases using JUnit. Deployed the application on WebLogic Application Server. Written Stored Procedures, Triggers, and Views extensively. Implemented the application and bug fixes in production environment. Documented related documents for future upgrades. Environment: Java 1.5, J2EE, Eclipse IDE, Servlets, JSP, Java Script, EJB, Java Beans, Spring, Hibernate, WebLogic 10.0, Rational Rose, JUnit, Log4J, JDBC, Oracle Confidential - Dallas, TX Java Developer Responsibilities: Involved in designing object model diagrams and data model diagrams to meet the requirements. Software development with agile (SCRUM) and TFD methodologies. Developed the application using different design patterns like Singleton, DAO and Session Façade. Designed the UI of the Docstore module with JavaScript, HTML, and CSS. Worked with Struts MVC for the client and the presentation layer of the application using JSP pages. Worked with Session beans (EJB) to create client interfaces for pricing, order submission, dynamic lead time, payment processing, and center availability. Developed complex SQL queries and stored procedures to add different media options to the database. Implemented web services (WSDL), UDDI in the business layer by providing the services to the UI from the external sources. Involved in JDBC Connection Pooling between J2EE and Oracle database (10g). Worked with JMS (MDB) for messaging. Responsibilities included setting up of queues, topics, troubleshooting various issues with messaging in dev/QA/production. Implemented integration of enterprise N-tier web based system using XML, Webservices (SOA). Used JUnit for unit testing different modules of the application. Wrote stored procedures to add new paper types and finishing options to the database. Developed the automatic build scripts using ANT for the application to deploy and test. Environment: WebSphere 6.1, EXT JS, HTML, SQL, JSP, CSS, UNIX, AJAX, Javascript, Oracle 10g, JDK 1.6Servlets 2.2, AJAX, JMS Confidential - Hartford, CT J2EE Developer Responsibilities: Analysis, design and development of functional components based on user requirements. Analysis, design and development of Application based on J2EE and Design Patterns. Developed front-end screens with JSP using Custom Tags and client-side validations using JavaScript. Developed the required Servlets. Implemented various design patterns viz. Front Controller, Session Facade, Business Delegate, etc. in Presentation and Business layer using Struts frameworks and session EJBs. Developed the EJB-Session Bean acts as Façade, will be able to access the business entities through their local home interfaces. Developed custom tag libraries for achieving most reusable code and ease of maintenance for presenting formatting and gathering data. Developed persistent classes using Hibernate mapping tool. Used Oracle as the backend database. Added new operations in the WSDL and their implementation to expose new services in the Web Services modules using Apache Axis tool. Created unit test cases using JUnit. Involved in writing various UNIX Shell Scripts used in the application. Compiled and built the application using Ant build tool. Deployed the application on WebSphere Application Server. Environment: Eclipse, EJB, JSP, HTML, XML, XSD, SOAP, WSDL, Apache Axis, Tag Libraries, Design Patterns, Servlets, Shell Scripts, Struts, Hibernate, JUnit, Oracle, Ant, IBM WebSphere Confidential - DURHAM, NC J2EE Developer Responsibilities: Analyzed the UI requirements and involved in the design documentation preparation and the development effort. Developed the JSP pages, Struts based controllers and configuration files. Developed the JMS modules for the integration of DE (Default Manager) with the back end host application. Developed PageFlows as part of TPLO(TouchPoint Loan origination) application development. Developed java applications for Java Database Connectivity. Worked on prototype development using hibernate2.3 for CL (Commercial Loans) Application. Developed JSP and Form Bean and Struts based controllers using Workshop 9.2 for CL Application. Created and executed Junit testcases as a part of regression testing. Environment: WebLogic Workshop 9.2, ALSB 2.5, UML, XML, WinCVS, Pageflows, JSP, JMS, SOAP, WSDL, WebServices, JDBC, Hibernate, TOAD VI, Unix/XP, Oracle 9i Confidential Software Engineer Responsibilities: Used Rational Rose for the Use Case Diagrams, Class Diagrams and Sequence Diagrams to represent the detailed design phase. Developed user interface using JSP 2.0 and HTML. Used Web services (SOAP) for transmission of large blocks of XML data over HTTP. Developed a web-based reporting for Credit Monitoring System with HTML, JSTL 1.2, custom tags. Used JMS for delivering confirmation messages to the businesses asynchronously. Used XML and SOAP with SAX parser to transfer data between applications. Used JAVA/J2EE Design patterns like Business Delegate, Session Facade, Data Transfer Object (DTO) and Service Locator in the project extensively, which facilitates clean distribution of roles and responsibilities across various layers of processing. Participated in database design using SQL Server. Used WebLogic Application Server 6.1 for deploying various components of application. Environment: BEA WebLogic App Server 6.1, Eclipse 2.1, Servlets, EJB, JDBC, JSP, JSTL, JMS, XML, SOAP, SAX, HTML, JavaScript, CVS, SQL Server 7.0	7	5	Male	['analytical', 'analysis', 'analysis', 'analysis', 'analysis', 'analysis', 'analyzed', 'confidential', 'confidential', 'confidential', 'confidential', 'confidential', 'lead', 'logic', 'logic', 'persistence', 'persistence', 'persistent', 'persistent']	['connection', 'connectivity', 'interpersonal', 'interpersonal', 'responsibilities', 'responsibilities', 'responsibilities', 'responsibilities', 'responsibilities', 'responsibilities', 'responsibilities', 'understanding']	19	12

Codebook

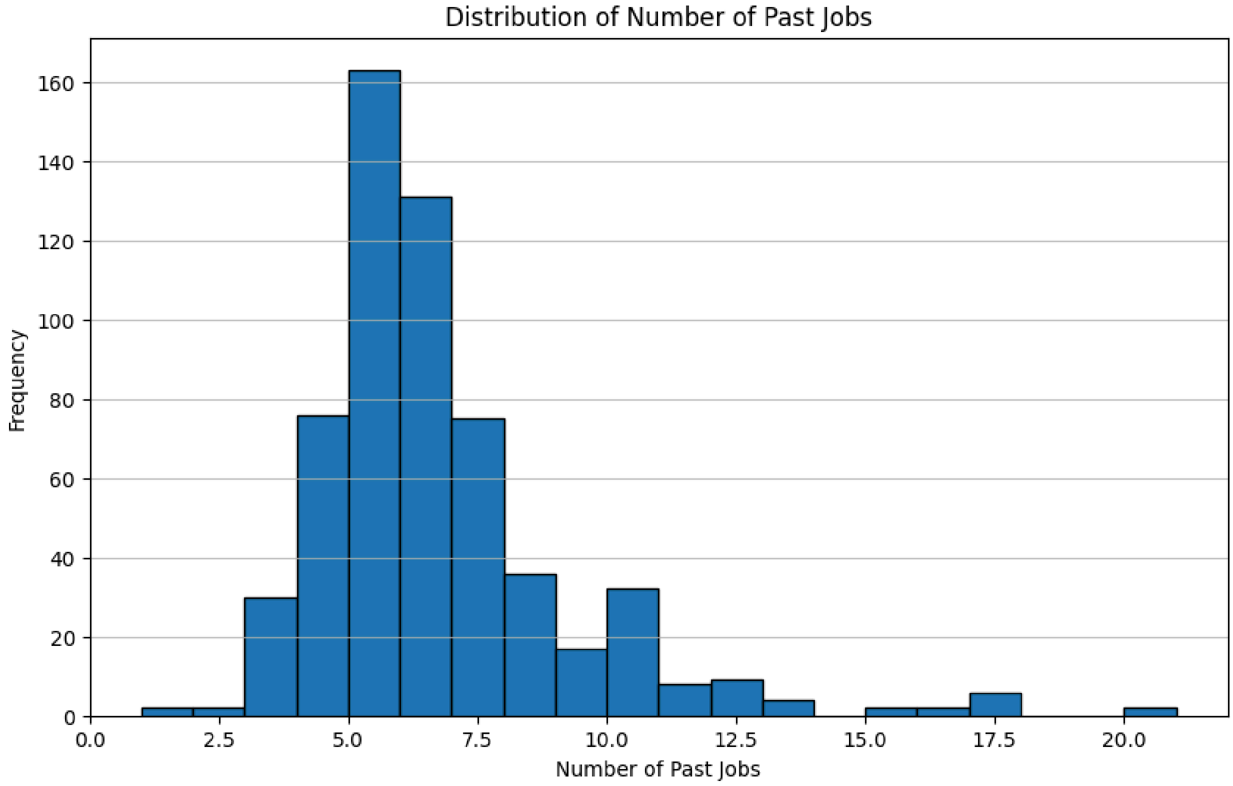
You must add a *codebook* – a description of all variables you are using, including ones you are creating for this project – to the README.md page of the data/ folder of your repo.

(OPTIONAL) Data Analysis

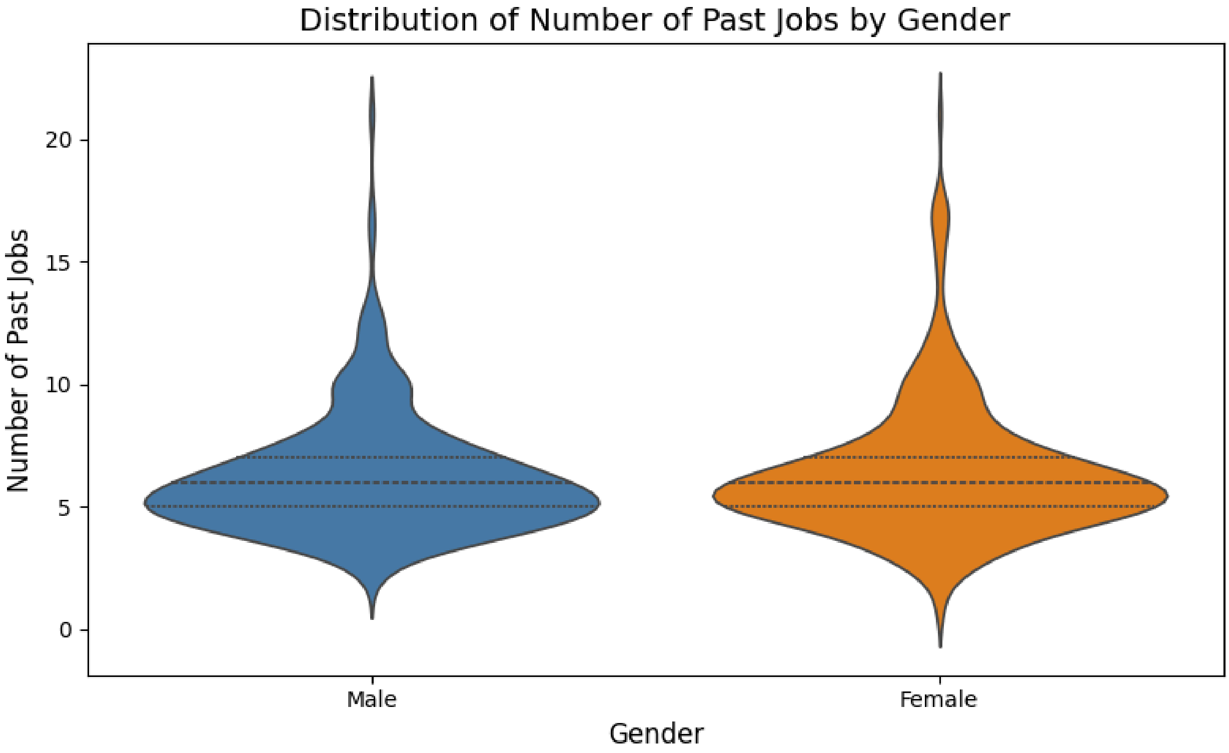
If you would like, you can start to sketch out some data analysis/modeling. This will not be counted for or against you, but I will give you feedback on it (which will be helpful for later).

Graphs And Analysis:

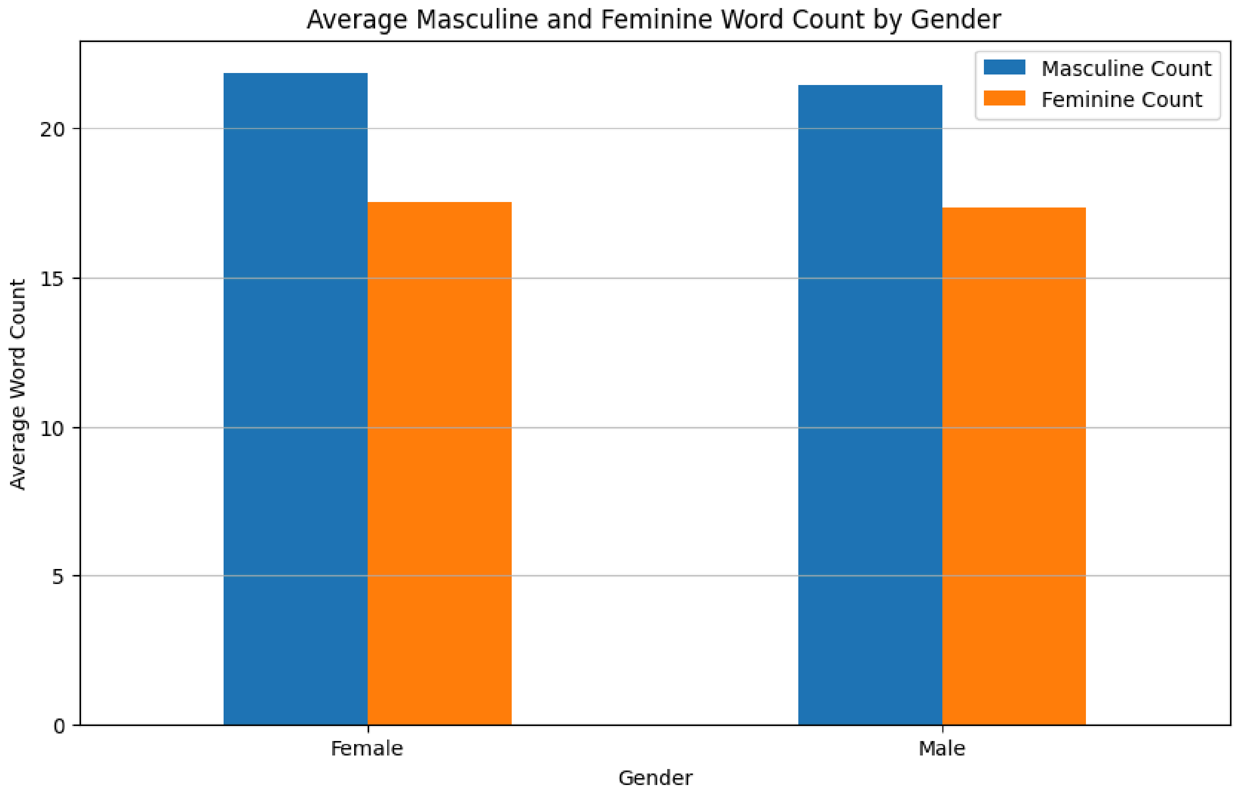
Below are the graphs that we created to visualize and explore our data.



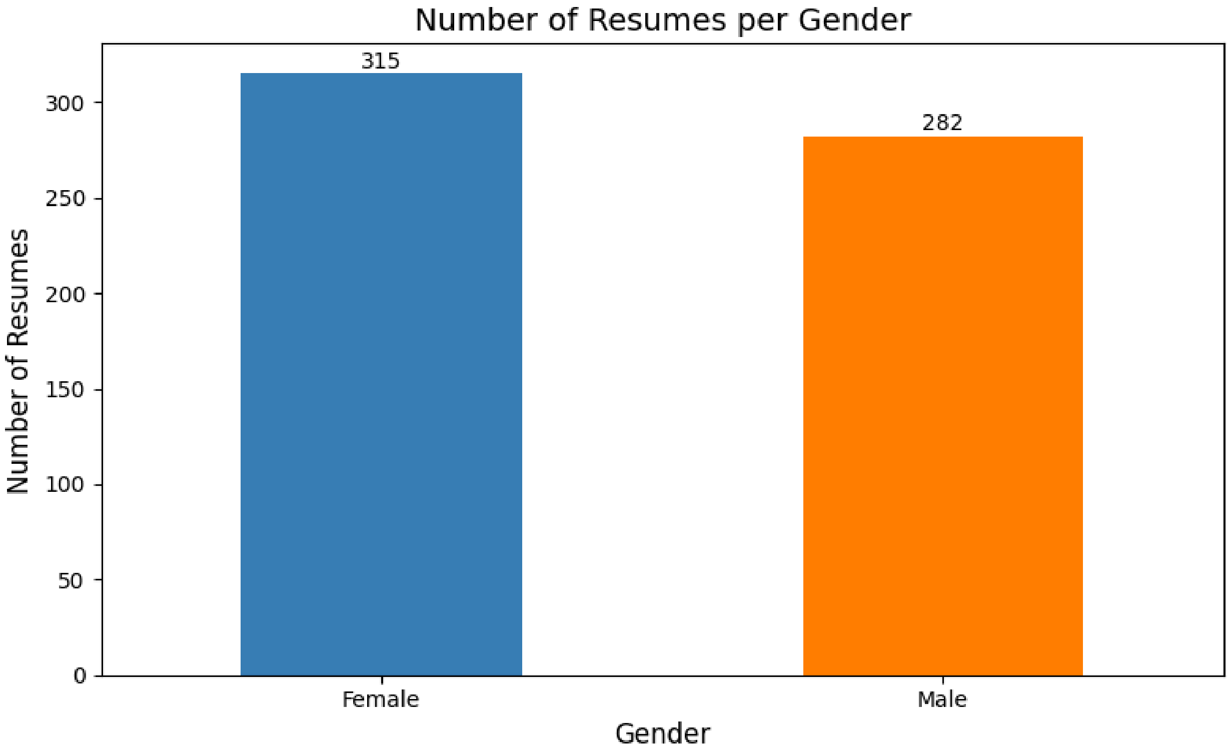
This visualization showed the distribution of number of past jobs listed in resumes in the dataset. The distribution is right-skewed and clustered around 5.0 to 7.5 years of experience.



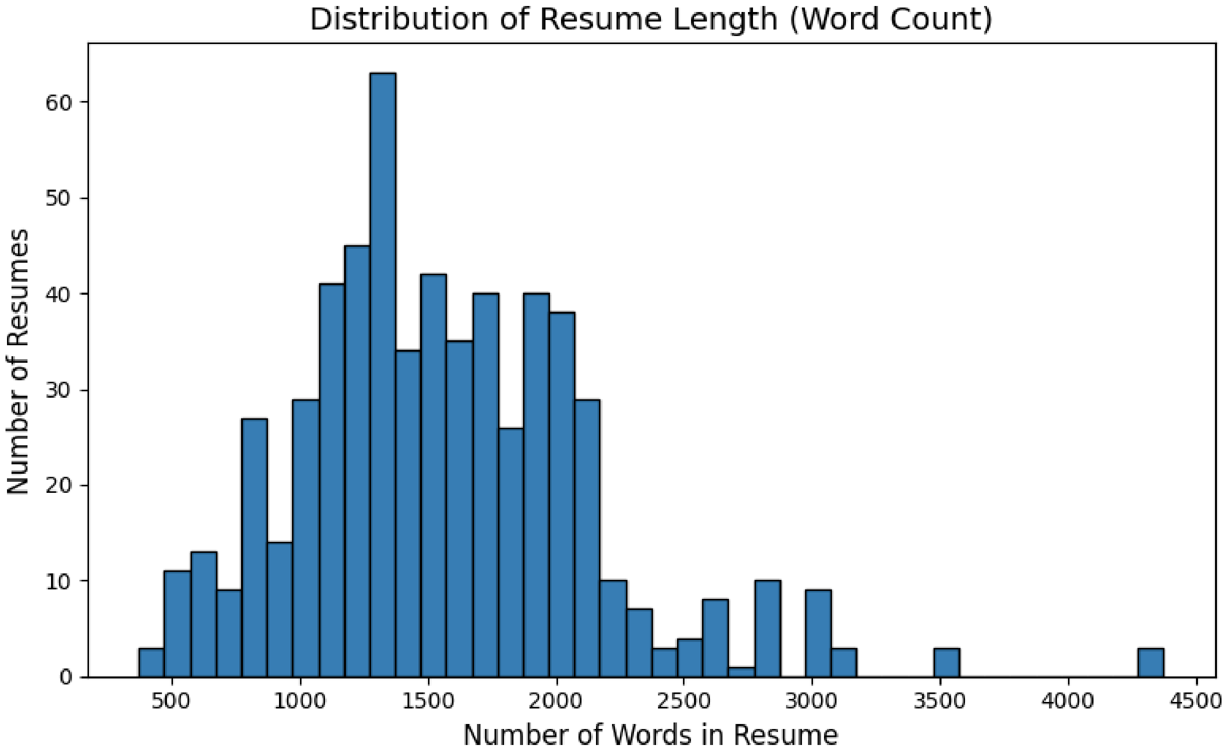
This visualization shows the distribution of the number of past jobs by gender. Our analysis shows that most applicants have held about 5 jobs, and the number of past jobs does not differ significantly by gender.



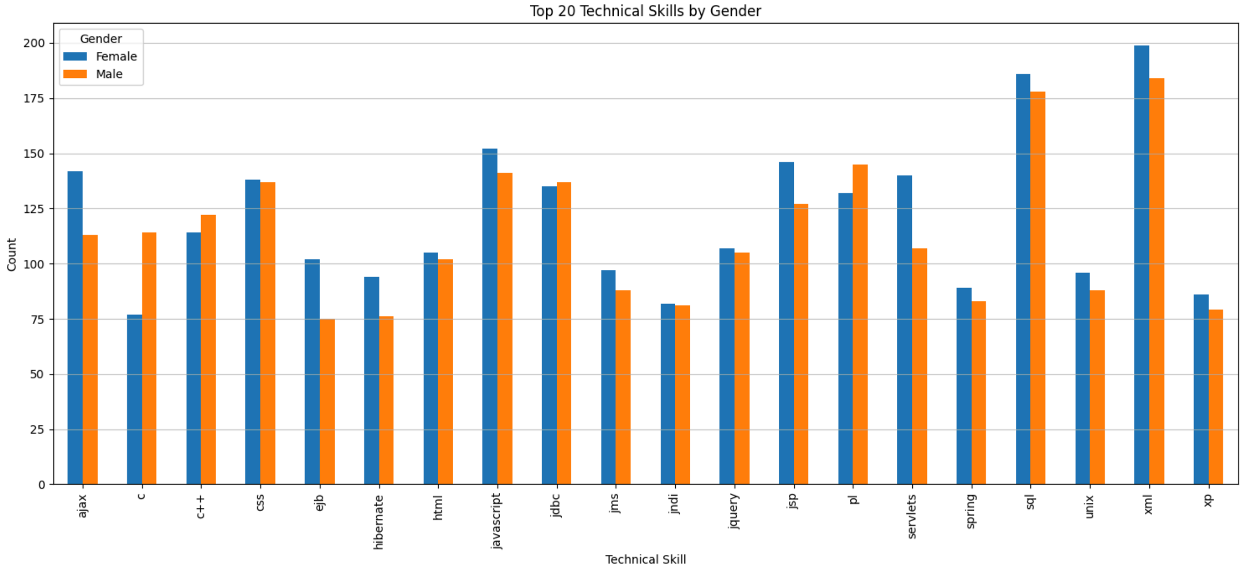
This analysis looked at the average number of masculine and feminine-coded words in resumes by gender. The results of this analysis show an equal number of gender coded words in both gender categories, with masculine-coded words being more prevalent than feminine coded words.



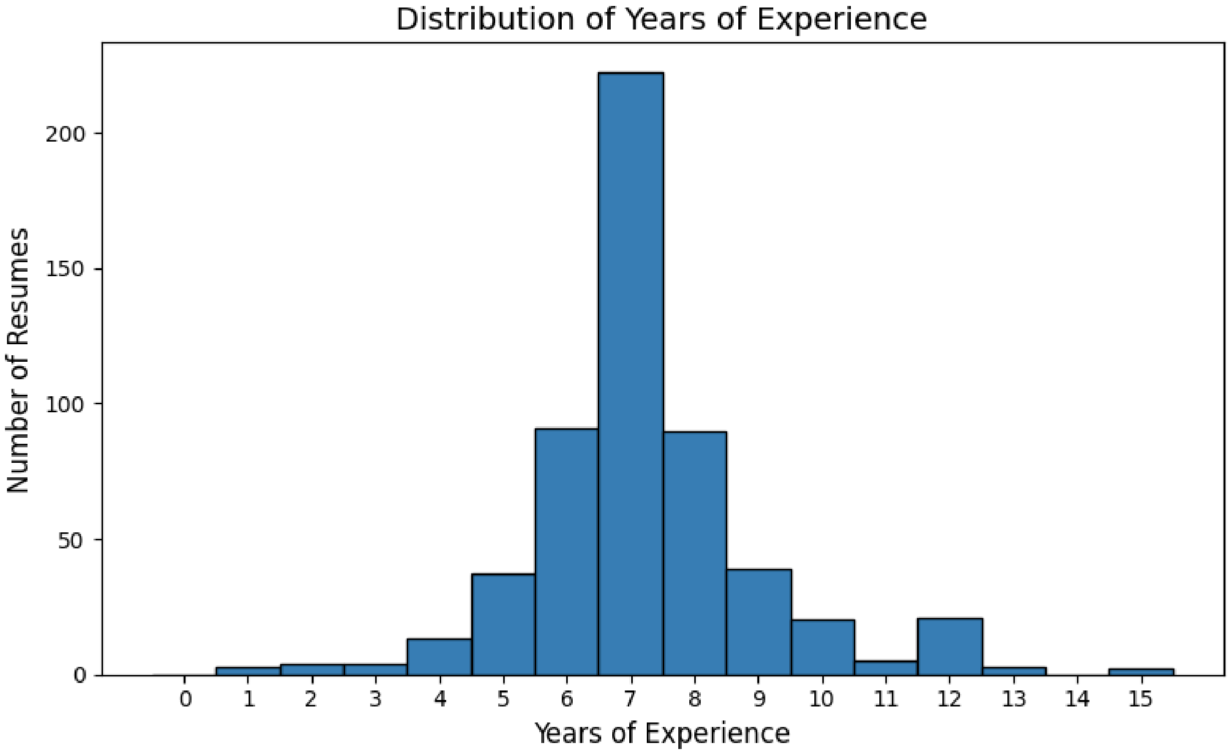
This analysis examined the number of resumes in the dataset by gender. There were 315 female resumes and 282 male resumes included in the dataset.

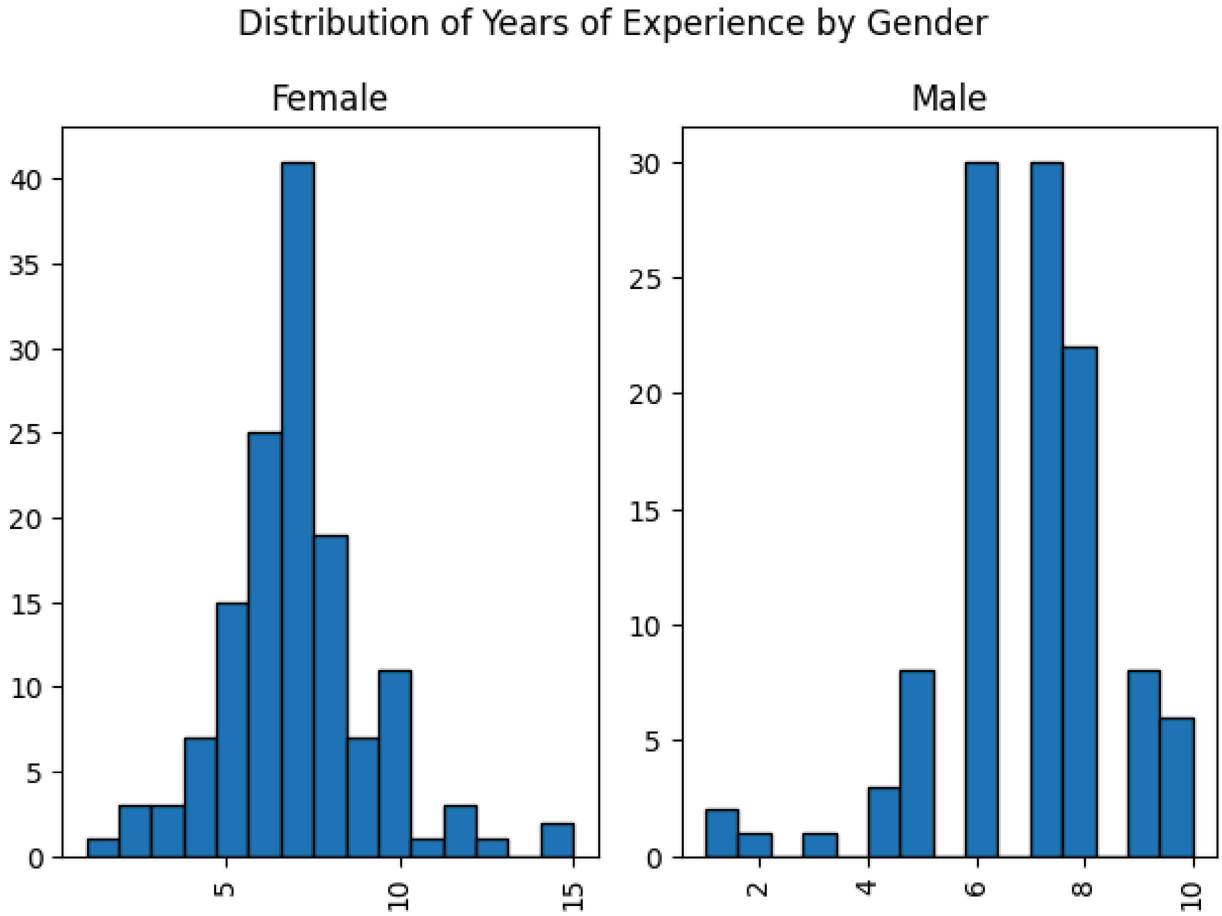


This visualization shows that most resumes have an average length of between 1000 and 2000 words. Very few resumes surpass 3000 words. Less than 10 resumes had below 500 words.



This analysis focused on the top 20 technical skills, with separate bars for female and male counts for each skill. In all skill categories except for C, C++, jdbc, and pl, there were more females than males with the skills.





This analysis focused on the distribution of years of experience, with an additional analysis showing the distribution of years of experience by gender. Overall, years of experience were normally distributed in the dataset, with the average number of years of experience for applications being 7.

Years of experience were more normally distributed for women than men, with a slight left skew for women and right skew for men. On average, men had more years of experience than women.

Formatting

All parts of the Exploratory Data Analysis should be professionally formatted. For example, this means labeling plots and figures, and using data preparation guidelines we went over in class (especially tidy and clean data we went over in Class 3). **The final product should be a .ipynb file. Both your original data file(s) and your cleaned data file(s) should be in your github repo, named accordingly.**

You should comment your code!

You must suppress all warnings and messages.

All plots must be professional in appearance, including meaningful axes and legend label and titles.

For the Exploratory Data Analysis assignment, you must display your code in the rendered output.

Please make sure that your code and plots are accessible. This means including alt-text for all plots and figures (this is in Description[]{} in LaTex)**, using color-blind friendly color palettes, and using patterns or textures that are noticeable in black-and-white! You will *lose* points for not doing this. This site and this site can help provide guidelines.**

Style and format do count for this assignment, so please take the time to make sure that everything looks good and that your data and code are properly formatted.

Repo organization

You should commit to your repo regularly as you work on your project, and you should keep your repo well organized.

Grading

Submit your Exploratory Data Analysis & Data Preparation assignment by **Wednesday, Oct 15 by 11:59 PM.**

The Exploratory Data Analysis & Data Preparation assignment will be graded as follows:

Total	50 pts
Exploring the data	10 pts
Data preparation and cleaning	10 pts
Codebook	5 pts
Data visualizations	20 pts
Formatting	4 pts
Repo organization	1 pt