DATA 301 Project 98%!!🎉

Milestone 1

Task 1: Form a team

Taii Hirano
Kokoro Imafuku
Yuki Isomura

Task 2: Choose a topic and a dataset, and get it approved

Source: Data Science Job Salaries

Task 3: Accept Project repository and understand the structure

Task 4: Create a project vision statement

The goal of this project is to provide a detailed dataset report about different jobs in the data science sector with its salaries for all students who plan to advance into the data science industry so that they can illustrate their future plans. We aim to accomplish this with various graphing technologies and python by the end of this course.

Questions:

- What is your project group number?
  Group 07

- Who are the members in your group? You will need to know their full names (hopefully you know more than their full names by this point).
  Taii Hirano, Imafuku Kokoro, Yuki Isomura

- Describe your dataset: In one sentence, describe your dataset.
  Dataset of different data science jobs with its salaries in different currencies, country of residence of employees, experience level, and remote ratio.

- What is the source of your dataset?: In addition to describing the source, please also include a link to where you found the dataset.

https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries

- License: what is the license of your dataset? CC0, MIT, etc…: If your license is not specified, you need permission to distribute it publicly in a GitHub repository.
  CC0: Public Domain

- Rows: How many rows of data are in your dataset? Combine across all files if you're working with multiple files.
  606 rows

- Columns: How many columns of data exist in your dataset? Combine across all files if you're working with multiple files.
  12 columns

- Interests: Why do you want to work with this dataset? What research questions/areas are you interested in? Note: Each member of the group should have their own research interests/angle/slant.
  Because it displays numerical and practical examples of data science students' future jobs.

  Taii: As a data science student, I'm interested in different kinds of jobs related to data science that I might pursue in the future. By analyzing this dataset, I would gain insight into the various salary levels in the data science field based on different countries. This creates a concrete vision of my future.

  Koko: One reason why I want to work on this dataset is because I would like to see different career paths that I could potentially advance into after graduation. Since the dataset offers important information like salary, job position name, etc it provides clear examples of my future.

  Yuki: One of my interests in this data set is the relativity of salary and the region they are working in. I believe this analysis helps me to decide where to work after graduating university. Another interest to this dataset is the profitability of remote working. Recent research shows that remote working is only beneficial for people who can work independently. Those people tend to earn more salary than other regular workers therefore I would like to analyze the correlation of salary and remote working.

Milestone 2

Task 1: Introduce and describe your dataset and topic

- Describe your dataset in about 150-200 words

  This dataset consists of various information in relation to different data science jobs including their work year, experience level, employment type, job title, salary in different currencies, employee residence, and the remote ratio. Our team has cited this dataset from Kaggle, and the data is collected from 2020 to 2022. It was collected by a dataset grandmaster on kaggle. The purpose of the data collection was not mentioned within the dataset.
  Our overall speculation of this dataset is that several factors, such as company location and the size of the company influence the salary level. Specifically, our team formed an assumption that certain locations of the company positively influence the salary level. We also conjecture that the company size and the salary level has a positive relationship. In contrast, we forecast that there are factors that have an indirect relation to salary such as remote working.

- Describe your topic/interest in this dataset - answer in about 150-200 words

  Our overall interest is the optimization of our job style for the future to earn a higher salary through analysis of salary data of workers in data science related fields. To break down this huge question into smaller pieces, we have formed our research questions based on our individual interest in this dataset.
  One of them is the effectiveness of remote working. Remote working is one of the huge advances produced by modern technology, but its effectiveness is still under research. Through this research, we may produce the relevance of salary and remote working, and moreover, acknowledgement about the skill we should gain before working in industries in order to be better adapted to remote working.
  Second one is the different salary levels based on the company size. We are interested in this because as data science students, we want to gain insight into different salary levels for each company size so that it will help me when choosing my career path.
  Lastly, our team would also like to see how salary levels differ in different countries. Although several factors contribute to how much a worker is paid, we would like to see which countries pay a worker more than the other. This information could benefit many data science major students, providing them with examples of career paths in different countries.

Task 2: Load your dataset from a file or URL

Task 3: Define and refine your research questions

Taii:

Which experience level worker is most hired in the mid-level company size?

Which region tends to have a large company size?
Which region tends to have a middle company size?
Which region tends to have a small company size?

Recent popularity in the data science/computer science field suggests the need for data analysis jobs. However, when we graduate from post-secondary education, we face the difficulties of choosing jobs. One of the reasons this problem occurs is that students do not have enough resources to determine their future careers. This problem could be solved by analyzing how the size of the company relates to the salary level. More in detail, we can analyze the relevance between regions and company size and the tendency of hired workers' experience levels in a specified scale of the company.

Yuki:

Does Remote Working Provide a Positive Impact on Their Salary?

Sub Questions:

- In what extent of workers' experience level, does remote working affect positively?
- In what type of jobs, does remote working affect positively?
- In which region, does remote working affect positively?

Due to the global pandemic, the benefit of remote working is recognized. However, recent research implies that successive rate and productivity with remote working depends on the skills of individual workers. In this research project, I will analyze the ratio of remote working, workers' level of skills, area of the profession, company size, and their location and compare them by salary based on the assumption that more salary means better success in their jobs. In general, salary is not an absolute criterion of success. But as there are a limited number of criteria in this dataset, I state the salary as a clear criterion of success in their jobs.

Kokoro:

What are the salary levels of workers in the United States and Canada?

In comparison to workers in these North American countries, how much do workers in other countries earn?
What are the salary levels of workers in other continents like Asia, Europe, and South America?

In recent years, the impression is that tech industries in North American countries, especially the United States, pay a higher salary to their workers than in other countries. I would like to deconstruct this idea and see if there is an existing relationship between the location of the company and salary in USD. I plan to do this by analyzing the company_location column, and the salary_in_usd column to see the different salary levels in different countries. I would categorize each country, and get the average salary of the workers in that country. I would also graph the different salaries in certain countries, and get both the high and the low-end salaries in that country. Furthermore, I would also analyze the different experience levels, to see if workers of a certain experience level are more likely to work in a certain company_location.

Milestone 3

Task 1: Conduct an Exploratory Data Analysis (EDA) on your dataset

As a rough guideline, each EDA should:

- Involve **at least two columns**/features of your dataset
- **At least three useful visualizations** created by you (the more the merrier (within reason)!)
- Some notes and commentary to help others see observations you find interesting.

Guide 1
Guide 2
Guide 3

Task 2: Refine your research questions

- You may have found that through your EDA you already answered your questions because they were more trivial than you thought, in this case you should come up with additional questions along the same track to demonstrate your proficiency with python, pandas, and seaborn (depending on your contracted grade).
- On the other hand, you may have found through your EDA that your research questions are impossible to address with your dataset due to limitations with the data, or for other logistical reasons (initially misinterpreting your dataset). In this case, you should pivot and hopefully your EDA showed you another potential path forward with a new research question or angle
    - This is fine and to be expected when trying to come up with research questions on a dataset you haven't worked with before!

Milestone 4

Task 1: Set up an “Analysis Pipeline”

Often when Data Scientists do analyses with the same or similar datasets, they set up an “analysis pipeline”. This has several advantages:

- record the steps so you can remember what you did.
- allows you to repeat the steps reproducibly, without doing a bunch of manual and repetitive work.
- make changes to the series of processing steps so you can improve and iterate.
- troubleshoot and debug errors in your processing.
- allows others to reproduce your analysis.
- if your data changes, you can update your outputs (report, images, etc…) easily without redoing all your processing.
- allows you to spend more effort and energy on your analysis and visualizations (if you do a good job with the pipeline).
- Here are some common steps of an analysis pipeline (the order isn’t set, and not all elements are necessary):

1. Load Data
    - Check file types and encodings.
    - Check delimiters (space, comma, tab).
    - Skip rows and columns as needed.

2. Clean Data
    - Remove columns not being used.
    - Deal with “incorrect” data.
    - Deal with missing data.

3. Process Data
    - Create any new columns needed that are combinations or aggregates of other columns (examples include weighted averages, categorizations, groups, etc…).
    - Find and replace operations (examples include replacing the string ‘Strongly Agree’ with the number 5).
    - Other substitutions as needed.
    - Deal with outliers.

4. Wrangle Data
    - Restructure data format (columns and rows).
    - Merge other data sources into your dataset.

5. Exploratory Data Analysis (not required for this Task).

6. Data Analysis (not required for this Task).
7. Export reports/data analyses and visualizations (not required for this Task).

For this Task, I will only ask you to set up a partial pipeline for the data loading, cleaning, processing, and wrangling steps.

Task 2: Method Chaining and writing Python programs

Step 1: Build and test your method chain(s)

Method chaining allows you to apply multiple processing steps to your dataframe in a fewer lines of code so it is more readable. You should avoid having too many methods in your chain, as the more you have in a single chain, the harder it is to debug or troubleshoot. I would target about 5 methods in a chain, though this is a flexible suggestion and you should do what makes your analysis the most readable and group your chains based on their purpose (e.g., loading/cleaning, processing, etc…).

Step 2: Wrap your method chain(s) in a function

```
def load_and_process(url_or_path_to_csv_file):
    # Method Chain 1 (Load data and deal with missing data)
    df1 = (
        pd.read_csv(url_or_path_to_csv_file)
        .rename(...)
        .dropna(...)
        # etc...
     )
    # Method Chain 2 (Create new columns, drop others, and do processing)
    df2 = (
        df1
        .assign(...)
     )
    # Make sure to return the latest dataframe
    return df2
```

Step 3: Move your function into a new .py file

- Inside your analysis/code/ directory each person in the group will create a project_functions.py file.
- Create a new file project_functions1.py, project_functions2.py, project_functions3.py (one for each student) and add the module imports you may need (pandas, numpy, etc…).
- Copy the load_and_process function into your project_functions.py file.
- Save the file.
- Add and commit it to your repository.
- Push it up to GitHub so that your teammates can also see this file.

- Each member of a group should now import the project_functions file in their analysis.ipynb file, and use the load_and_process function
- A Jupyter Notebook cell should look something like (with the appropriate relative import):

Task 3: Conduct your analysis to help answer your research question(s)

Milestone 5

Task 1: Process your data for your Tableau Dashboard

You should prepare and process your data so that when you create your dashboard, you have to do minimal data wrangling or manipulation in Tableau. There are many different ways to deal with this, but I suggest you export a dataset that you can easily use in Tableau to plot whatever you need to in your dashboard.

You should put the exported .csv files in the data/processed directory.

Task 2: Create a Dashboard using Tableau

Create a dashboard (using your processed dataset - i.e. you do not have to do the data cleaning, wrangling, processing again) to create a Dashboard using Tableau. There are no requirements for this Dashboard, but please remember that you will be graded based on the quality of your dashboard, and how well it answers your research questions and/or helps with the exploratory data analysis. I suggest you take this opportunity to explore as many of the features that make sense for your project, and get help from us when you need it! You should place the Tableau file in the dashboard folder.

Each person in the group should have their own Tableau Dashboard, but if you can find a way to combine it into one dashboard, that's also fine. I suggest using multiple "tabs" in Tableau to split up research questions or parts of the dashboard.

Task 3: Present your dashboard

For this Task, you will record a video (Explainer video) showing your Tableau Dashboard.

Groups of 3 project presentations should be 7-10 minutes long.

Other requirements:

All members of a group must participate in the Explainer Video (you may have multiple tabs in your Tableau Dashboard).

Your recording does NOT have to have a high production value (editing, background sound, video titles etc..) and I would suggest not spending too much time on the non-dashboard component.

Project Dashboard Presentations

It is very important for you to know that we are NOT looking for hollywood production value here. We want to see your Tableau dashboard, get a walk-through of the key features, and hopefully see some enthusiasm about your project.

Here is how you should allocate your efforts for these short videos:

[10%] : Timely submission of a video link or file by the deadline and before the grace period ends.

[20%] : Clear explanation of the project research questions and information about the dataset.

[50%] : Guided walk-through of the key features of the Tableau Dashboard.

[30%] : Show how your dashboard data answer your research questions.

Recording your Dashboard Presentation

Here is a video that you can watch to go through the entire process of creating an Explainer Video. I highly recommend you to use Zoom's screenshare and record functionality as it's by far the easiest way to record your video.

Taii Script:

Before we dive into the analysis, let me give you an overview of our dataset. We collected data on salaries for data science related jobs, including information on year, experience level, salary, employee residence, remote ratio, company location, company size and so on. We set salary as our measure of success for this project, and did not consider factors such as working hours or overall job satisfaction.

One important note is that the salary currency is reported in real-time exchange rates, which may not be entirely accurate. Additionally, we acknowledge that there may be more appropriate ways to categorize regions than how we have currently defined them.

One more thing, we attempted to create violin plots to visualize the data, but encountered difficulties in creating them using Tableau so we used box plots instead. With that said, let's move on to my analysis of the best company size in North America for each experience level based on their salary.

To answer my research question, I have divided it into three sub-questions.

Firstly, I will explore how the company size is distributed across the region.

- We found that most of our data was collected from companies in North America, where companies with medium size were almost twice as common as those with large size.
- We can see the similar distribution in Europe.
- On the other hand, in the Asian region, there were more large size companies than medium size companies.
- For other regions, such as Latin America and the Middle East, there appear to be more small-sized companies. However, due to the relatively smaller number of samples from these regions, it is not entirely reliable to conclude that they tend to have a higher proportion of small-sized companies.

In the next dashboard, I will investigate the distribution of experience levels for each company size. This barplot displays the experience level of workers hired in each company size.

- Overall, medium size companies have the largest number of employees.
- In large size companies, mid-level experience workers are predominantly hired, though the population of workers with senior-level experience is also significant. This could be due to the fact that larger companies require a larger number of mid-level and senior-level workers to handle bigger projects, with fewer expert-level workers.
- For medium size companies, it is clear that senior-level experience workers are the most commonly hired among the four experience levels. I speculate that this is because medium companies are limited in their hiring capacity while working on relatively larger projects, thus requiring senior-level workers who can work independently.
- In small size companies, the numbers of workers with entry-level and mid-level experience are almost equal and the highest. This could be because small companies may not require many workers with senior or expert-level experience, as they tend to handle smaller projects.

Finally, I will focus on the North America region since our data had a large number of samples from this region.

Analyzing the salary distribution across experience levels and company sizes,

- I observed that as the experience level becomes more advanced, the salary level increases, with the most common salary range for entry, middle, and senior level employees being between $100,000 and $200,000 US Dollars.

- Additionally, we found that salaries tend to increase as the company size gets bigger for these groups.
- However, we also found that expert-level employees had a larger salary range compared to other experience levels, especially in small size companies.
- This suggests that they may consider starting their own business as they gain skills, which allows them to work in a small company while earning a lot.

In conclusion, I recommend medium size companies in North America for entry, middle, and senior level employees, while expert-level employees can explore the option of starting their own business for the potential to earn more.

Milestone 6

Task 1: Address project feedback

Your assigned project TA should have created an issue in your repository with some feedback for you to address on your milestone. You should address this feedback to the best of your ability.

Task 2: Final Report: Create a single markdown file that will be your final report

- Once you're done your analysis and you've addressed all the TA and instructor feedback to improve your project, you will export your final figures as PNG (or JPEG) files, and then consolidate your findings into a single markdown file.
- The name of your markdown file should be: final_report_groupYY.md (replace YY with your group number), and it should be located in the root of your project repository.
- This markdown file is a major deliverable and will require some coordination amongst your teammates to ensure that it is a cohesive and complete document, that provides a summary of all your hard work.
- This markdown file should NOT have any code in it, it is meant to be a narrative/summary of your exploratory data analysis, as well as your actual analyses.
- You should of course, link to your jupyter notebooks because a portion of your audience will be interested in digging deeper into the analysis and looking at the code.
    - For example, this can look like: "You can find the full analysis notebook here, including the code and the data here
- Here are the suggested sections of your Final Report:

- Introduction: A short paragraph introducing your project to the audience and a motivation for why this project is important. It's fine to say your group has an interest in this topic and were keen to explore it more.
- Exploratory Data Analysis: A summary of the highlights of your EDA, where you can show some visualizations of the exploratory data analysis your group did.
- Question 1 + Results: Clearly state your research question, and include 2-3 visualizations that helped you answer your research question. You can create multi-panel figures, but each of your visualizations must speak directly to your research question, and any insights you were able to get from it should be clearly articulated in the figure caption/description.
- Question 2 + Results: Same as above.
- Question 3 + Results: Same as above.
- Summary/Conclusion: A brief paragraph that highlights your key results and what you learned from doing this project.

Task 3: Make your repository public

1. On GitHub, navigate to the main page of the repository.
2. Under your repository name, click Settings.
3. Under "Danger Zone", to the right of to "Change repository visibility", click Change visibility.
4. Select a visibility.

Task 4: Create a release for your repository

Create a new release for your project called 0.1.0.

As you progress through your project, you can add more releases, once you're happy with where you're at, I suggest creating a new release at 1.0 (perhaps just before you're submitting your final milestone?).

Creating a new release

Here are the instructions to create a release.

Briefly, here is how to create a release (from the docs):

1. On GitHub, navigate to the main page of the repository.
2. To the right of the list of files, click Releases or Latest release.
3. Click Draft a new release.
4. Type a version number for your release. Versions are based on Git tags.

5. Use the drop-down menu to select the branch that contains the project you want to release.
6. Type a title and description for your release.
7. Once you're ready to publicize your release, click Publish release. To work on the release later, click Save draft.

Final Report - Group 07

Introduction

Our project aims to analyze optimized job styles for the future to earn a higher salary through the analysis of salary data of workers in data science related fields. With the increasing demand for data science professionals, it's essential to understand the factors that influence salary levels, such as remote working, company size, and location. Our team believes that this project is important as it can provide valuable insights into career paths and help students make informed decisions about their future. We are excited to explore the data and test our hypotheses about the various factors that influence salary levels in the data science field. Our project is based on a dataset from Kaggle, collected from 2020 to 2022, and includes information on work experience, employment type, job title, salary, employee residence, and remote ratio. We hope that our findings will contribute to the overall understanding of the data science job market and inform future career decisions.

Exploratory Data Analysis
For each research question, we have three different highlights.

- We utilized violinplot to illustrate the correlation among salary, experience level, and company size, enabling us to accurately display the distribution of salaries for workers at each experience level.
- For the second research question, we utilized barplots to effectively illustrate the varying salary levels across different continents. Additionally, we incorporated a pie chart to indicate the distribution of data among different countries and to demonstrate the dataset's strong bias towards the United States.
- To visualize the salary distribution for different experience levels of workers by ratio of remote working, we have used a box plot so that the more precise tendency in effectiveness of remote working for each level is better readable.

Research Question 1 and results: What is the best company size in North America for each experience level based on their salary?

To address this research question, I divided it into three sub-questions.

1.1: Which experience level workers are hired in each company size?

Firstly, I investigated the distribution of experience levels for each company size.

The results showed that medium-sized companies had the largest number of employees overall. In large-sized companies, mid-level experience workers were most commonly hired, followed by senior-level workers. For medium-sized companies, senior-level workers were the most commonly hired. Small-sized companies tended to hire entry-level and mid-level workers in almost equal numbers.

1.2 How are the each company size distributed over the region and which level of company size do they tends to have?

Next, I explored the distribution of company size across different regions.


The data showed that in North America and Europe, medium-sized companies were the most common, while in Asia, large-sized companies were more prevalent. In other regions, such as Latin America and the Middle East, there appeared to be more small-sized companies.

1.3 What is the difference in salary level by experience level for each company sizes in North America?

Finally, I focused on the North American region and analyzed the salary distribution across experience levels and company sizes.

The above plot revealed that as the experience level increased, so did the salary range, with the most common range for entry, middle, and senior-level employees being between $100,000 and $200,000 US Dollars. Moreover, salaries tended to increase as the company size grew, except for expert-level employees who had a wider salary range, particularly in small-sized companies.

In conclusion, based on the results, I recommend medium-sized companies in North America for entry, middle, and senior-level employees, while expert-level employees may consider starting their own business for the potential to earn more.

Research Question 2 and Results: What are the salary levels of workers in the United States and Canada?

I have added few sub questions in relation to my main research question:

- In comparison to workers in these North American countries, how much do workers in other countries earn?

- What are the salary levels of workers in other continents like Asia, Europe, Latin America, and Africa?
- How much does workers in different regions earn on average if they are working: Completely remote, Partially remote, and Not remote?

The analysis of the salary data from the dataset revealed interesting insights into the salary levels of workers in different countries, including the United States and Canada. The average salary level in United States was approximately $144,000. In contrast, the average salary level in Canada was approximately $100,000. Thus, the difference in salary level was approximately $44,000.

In comparison to the two North American countries above, this graph showed that workers in Russia earn the highest salaries, amounting to approximately $160,000. The second highest was the United States, followed by New Zealand at $125,000. Overall, European countries tended to be in the mid-level range, while Asian countries were distributed equally throughout the graph. One key point about this observation is that Japan and Singapore are on the highend of salary level, but countries like Vietnam and Pakistan are on the lowend. Therefore, this observation can indicate how third-world countries tend to pay their workers less than first-world countries. This indication is further demonstrated by the salary levels in Latin American countries compared to Oceanian countries like New Zealand and Australia. However, it's worth noting that the dataset is heavily skewed towards the United States, with 62% of the data coming from the US. Despite categorizing all countries with 5 or less data as "Others", this still counted for only 7 percent of the entire dataset.

As for the sub-question of the effect of remote work on salary levels in different regions, the plot showed that the impact varies by region.

In Europe and North America, a combination of in-person and remote work results in the lowest salary levels, while in Asia, both in-person and partially remote work result in better salary levels than full remote work. The limited data from other regions makes it challenging to draw firm conclusions on this topic.

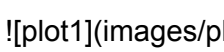
In conclusion, with the assumption that this dataset is heavily skewed towards the United States, it can be said that United States offers the highest average salary levels, and therefore, would be recommended for other data science students to find a job in the United States.

Research Question 3: Does Remote Working Provide a Positive Impact on the Salary of Employees?

The COVID-19 pandemic has brought attention to the advantages of remote work. However, recent studies suggest that the effectiveness and productivity of remote work largely depend on individual workers' skills.
In this research project, I analyzed the correlation between the ratio of remote work and the salary and in which condition it works more effectively. This analysis is under the assumption that a higher salary indicates better job performance. Although salary alone is not an

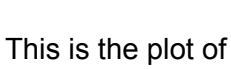
absolute measure of success, as there are limited criteria in this dataset, I used it as a clear indicator of job success.

3.1 Correlation of Remote Working and Salary in General



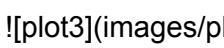
This is the plot of salary distribution by ratio of remote working.
Referring to the plot, utilizing both remote and in-person working methods appears to be the least favorable option. This observation is mostly consistent throughout the analysis with additional conditions.
There also is not a significant contrast between solely working in-person or remotely.

3.2 Workers' Experience Level



This is the plot of salary distribution for different experience levels of workers by ratio of remote working.
To begin with, using a combination of in-person and remote working resulted in the lowest productivity regardless of experience level, as previously observed in plot in *3.1*.
Furthermore, as experience level increases, the salary gap between partial remote and full remote becomes more significant.
However, for Executive-level experienced workers, the disparity between working fully in-person or fully remote is noticeable, indicating a clear advantage to remote work.

3.3 Type of Jobs



This is the plot of salary distribution for different jobs by ratio of remote working.
To avoid insufficient data, I have selected several jobs with more than 10 workers to compare them through plots since it is challenging to categorize jobs without knowing their precise duties.
Based on the plot, the effectiveness of remote work varies by job. Jobs such as Data Scientist, Data Engineer, and Data Analyst are similar, and their distributions are alike. These three jobs also represent 61.29% of the total data, which results in a homogeneous shape in the plot similar to plot in *3.1*.
However, Machine Learning Engineer, Research Scientist, and Data Science Manager exhibit a property where full in-person work provides better advantages than full remote work, which is inconsistent with *3.1*. Nonetheless, it's crucial to note that these jobs have relatively less data.
In the case of Data Architect, there is a lack of data, and all workers belong to full remote work, making it impossible to evaluate its effectiveness.

Overall, the efficacy of remote work varies among jobs, but it's difficult to identify similarities due to limited data sources.

3.4 Region

This is the plot of salary distribution for different regions by ratio of remote working.
According to the plot, the effectiveness of remote work appears to be unique to the region. In Europe and North America, the shape of the plots is very similar to plot in *3.1*. The combined use of in-person and remote work yields the worst productivity, and full in-person and full remote work exhibit negligible differences. However, this occurs because the sum of them represents 89.29% of the total data.
Conversely, in Asia, full in-person work and the combined use of in-person and remote work are considerably more effective than full remote work. The cause of this phenomenon cannot be analyzed, but it is an intriguing property worth noting.
In the other remaining regions, the data is not accurate enough to analyze.

3.5 Transition Over the Years

This is the plot of the transition of average salary over the years.
This plot reveals a consistent trend in which the average salary of data scientists increases over the years, regardless of the remote working ratio. There are three major factors that could impact the average salary of data scientists.

- The COVID-19 pandemic caused significant damage to the economy in 2020, but it has gradually recovered since then.
- Data science is a profession that may not necessarily require in-person work, which could mean that the average salary of data scientists is less negatively affected than other professions.
- The economy in general tends to grow gradually, meaning that the average salary of data scientists should increase even without completely recovering from the COVID-19 damage.

This is the plot of the transition of average salary over the years by ratio of remote working.

Following precedent, the combination of in-person and remote work tends to result in the worst productivity, as shown in the analysis *3.1*.
Based on the plot, it appears that the average salary of full in-person jobs decreased in 2021, which is an intriguing finding. It is possible that this was caused by some companies going out of business that year. In contrast, companies that had adopted remote working may have been able to avoid bankruptcy. However, due to the limited range of data available

(only from 2020 to 2022), it is difficult to draw a definitive conclusion about the underlying causes of this phenomenon.

3.6 Same and Different Locations as Company

These are the plot of salary distribution by ratio of remote working when employees are living in the same and different countries as the company.
The plots suggest that remote working is more beneficial for employees residing in different countries than those residing in the same country as their company. The group of first plot accounts for 91.6% of the data, showing that there is not much difference between full in-person and remote working when employees have an option to work in-person. However, for employees without an option to work in-person, more remote working appears to result in better productivity. Due to limited data in the group of second plot, the reliability of the second plot may be lower than the first plot.

Summary
Following are the summary of the analysis about effectiveness of remote working under various conditions.

- Partial remote working is generally the worst option.
- Experienced workers tend to benefit more from remote working.
- Remote working can increase the salary of employees if they live in a different country than their company.
- Effectiveness of remote working varies by job.

Conclusion

In conclusion, based on the analysis, medium-sized companies in North America appear to be a favorable option for entry, middle, and senior-level employees. However, expert-level employees may want to consider starting their own business for the potential to earn more. Additionally, the dataset heavily favors the United States, which offers the highest average salary levels, making it an attractive destination for data science students to find a job. As for remote working, the effectiveness varies by job, and experienced workers tend to benefit more. Partial remote working is generally the least favorable option. Interestingly, remote working can also increase an employee's salary if they live in a different country than their company. Overall, individuals should consider their specific situation and preferences when deciding on the best working arrangement for them. As for all the university students in the data science field, on the premise that these students are actively seeking for a job in this industry, and they are low-skilled data scientists, we suggest to enter the job market in the United States at a mid-sized companies and to avoid partially remote jobs.