- ## MILESTONE 1: Define Problem/Problem Understanding

Activity 1: Specify the business problem;

The goal of this project to predict doctors' Annual salaries using machine learning is to develop a model that can accurately estimate a doctor's salary based on various factors such as their education level, specialty, years of experience, location, and other relevant variables.

Activity 2: Business requirements;

1. Collect and segregate the salary data for various doctors.

2. Perform exploratory data analysis to identify key trends and patterns in the salaries.

3. Develop data visualizations using various graphs and charts for better understanding of the predictions and clarity.

4. It should show real time predictions so that we understand the needs of various doctors and their expectations.

Activity 3: Literature Survey;

1. Using machine learning to predict physician income: A case study in Canada" by Jean-Francois Ethier et al. (2020): This study developed a machine learning model to predict physician income based on factors such as age, gender, education, and practice characteristics. The study found that the model had an accuracy rate of 72%, which
2. Predicting physician compensation: Machine learning and physician practice characteristics" by John D. Gazewood et al. (2020): This study used machine learning algorithms to predict physician compensation based on practice characteristics such as patient volume, patient demographics, and practice location. The study found that the model had an accuracy rate of 87%, indicating that machine learning can be a useful tool in predicting physician compensation.

Activity 4: Social or Business Impact;

1. **Social impact**: From a social perspective, accurate salary predictions can help ensure that physicians are fairly compensated for their work. By identifying the factors that influence salaries, machine learning models can help address potential biases and disparities in compensation, particularly regarding demographic factors such as gender and ethnicity.
2. **Business impact**: From a business perspective, salary predictions can help healthcare organizations better manage their resources and allocate compensation in a more strategic manner. By identifying the factors that influence physician salaries, organizations can develop more effective compensation structures that attract and retain top talent while managing costs.

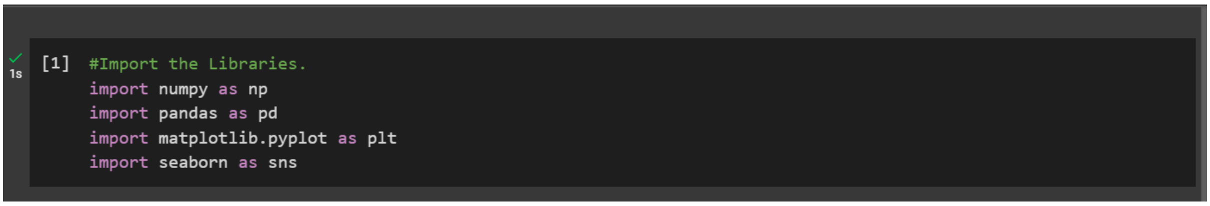
MILESTONE 2: Data Collection & Preparation;

Activity 1: Collect the Dataset;

https://data.world/aik/u-s-doctors-pay-2016

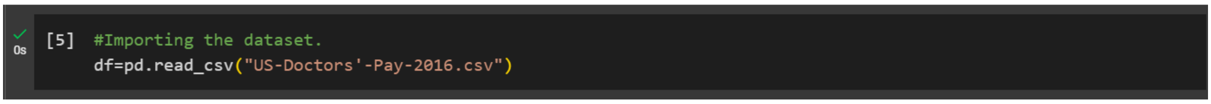
Download and upload the Dataset in Google Collab/Jupyter Notebook.

Activity 1.1: Importing the libraries;



```
[1] #Import the Libraries.
    import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
```

Activity 1.2: Read the dataset;

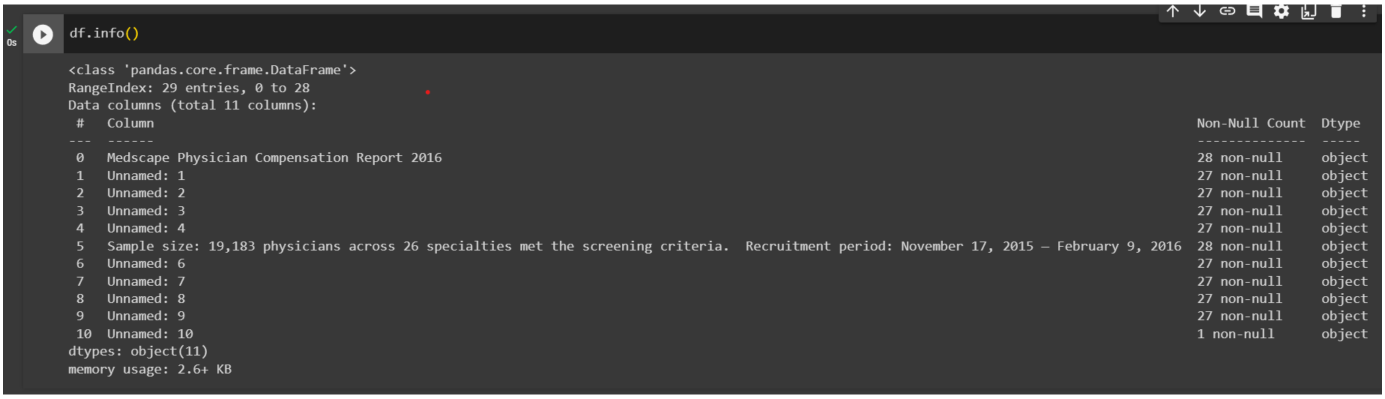


```
[5] #Importing the dataset.
    df=pd.read_csv("US-Doctors'-Pay-2016.csv")
```

Activity 2: Data Preparation;

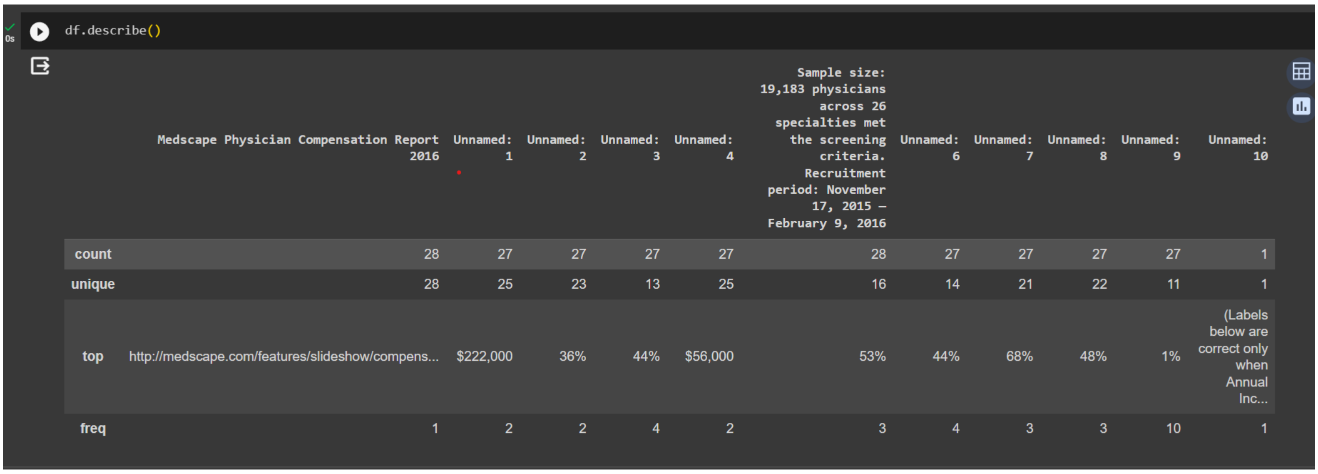
The values in the downloaded dataset is filled with random values which cannot be trained and hence we need to filter it in order to achieve good results.

Activity 2.1: Handling missing values;



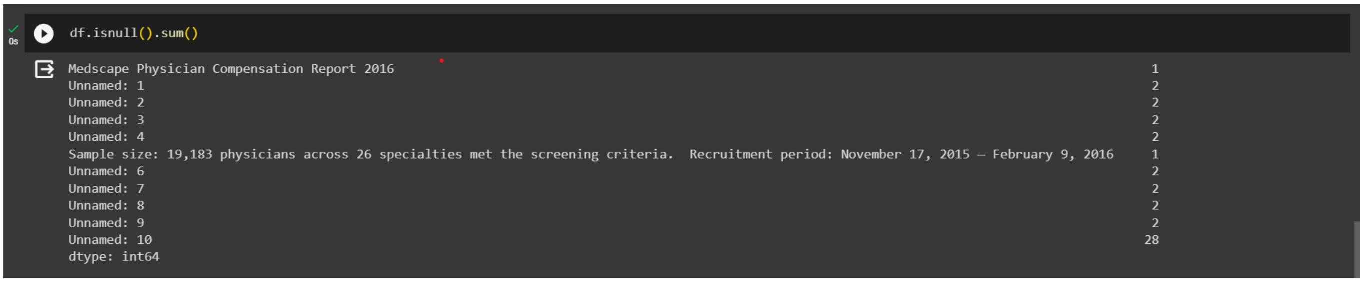
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29 entries, 0 to 28
Data columns (total 11 columns):
 #   Column                                                                                                                    Non-Null Count  Dtype 
---  ------                                                                                                                    --------------  ----- 
 0   Medscape Physician Compensation Report 2016                                                                               28 non-null     object
 1   Unnamed: 1                                                                                                                27 non-null     object
 2   Unnamed: 2                                                                                                                27 non-null     object
 3   Unnamed: 3                                                                                                                27 non-null     object
 4   Unnamed: 4                                                                                                                27 non-null     object
 5   Sample size: 19,183 physicians across 26 specialties met the screening criteria.  Recruitment period: November 17, 2015 – February 9, 2016  28 non-null     object
 6   Unnamed: 6                                                                                                                27 non-null     object
 7   Unnamed: 7                                                                                                                27 non-null     object
 8   Unnamed: 8                                                                                                                27 non-null     object
 9   Unnamed: 9                                                                                                                27 non-null     object
 10  Unnamed: 10                                                                                                               1 non-null      object
dtypes: object(11)
memory usage: 2.6+ KB
```



```
df.describe()
```

	Medscape Physician Compensation Report 2016	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Sample size: 19,183 physicians across 26 specialties met the screening criteria. Recruitment period: November 17, 2015 – February 9, 2016	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10
count	28	27	27	27	27	28	27	27	27	27	1
unique	28	25	23	13	25	16	14	21	22	11	1
top	http://medscape.com/features/slideshow/compens...	$222,000	36%	44%	$56,000	53%	44%	68%	48%	1%	(Labels below are correct only when Annual Inc...
freq	1	2	2	4	2	3	4	3	3	10	1

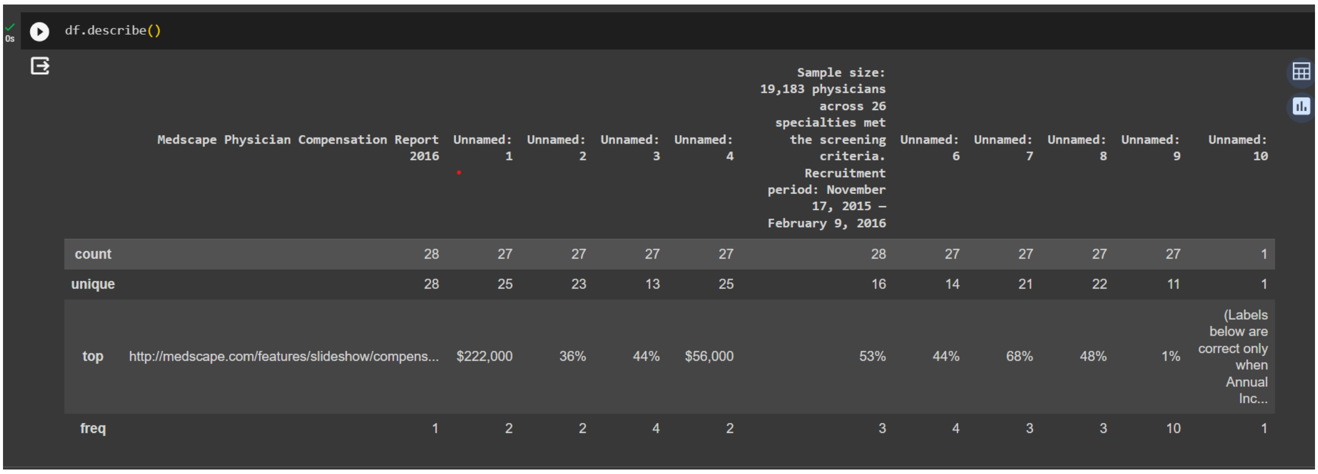


```
df.isnull().sum()
```

```
Medscape Physician Compensation Report 2016                                                                                 1
Unnamed: 1                                                                                                                  2
Unnamed: 2                                                                                                                  2
Unnamed: 3                                                                                                                  2
Unnamed: 4                                                                                                                  2
Sample size: 19,183 physicians across 26 specialties met the screening criteria.  Recruitment period: November 17, 2015 – February 9, 2016   1
Unnamed: 6                                                                                                                  2
Unnamed: 7                                                                                                                  2
Unnamed: 8                                                                                                                  2
Unnamed: 9                                                                                                                  2
Unnamed: 10                                                                                                                28
dtype: int64
```

- ## MILESTONE 3: Exploratory data analysis;
  ### Activity 1: Descriptive statistical;

  Descriptive analysis is to study the basic features of data with the statistical process. Here pandas have a worthy function called describe. With this describe function we can understand the unique, top, and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features.

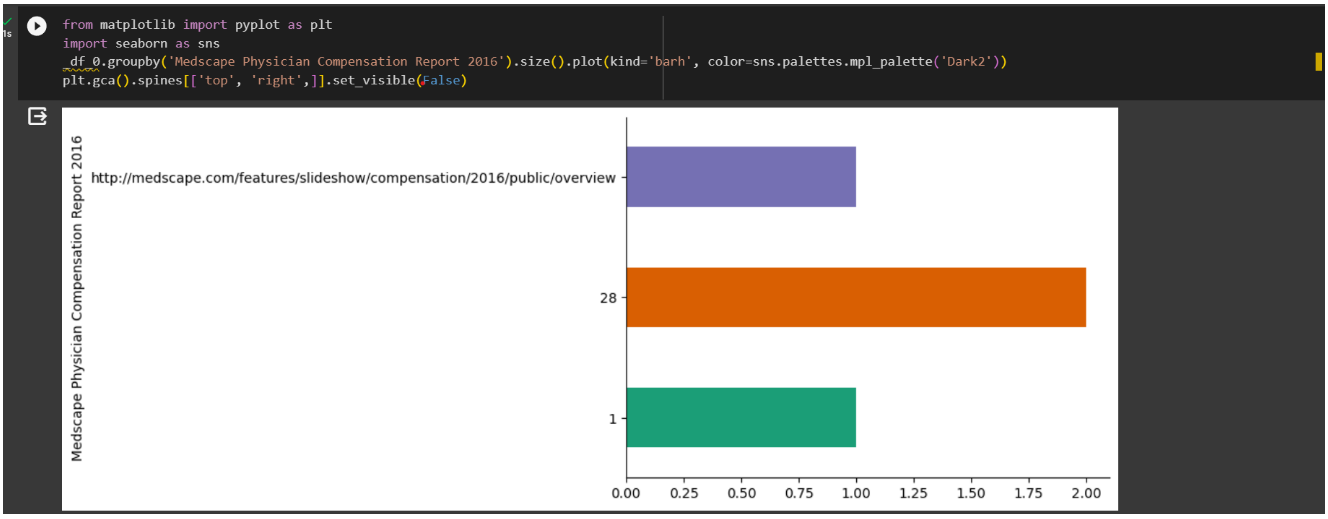


```
df.describe()
```

	Medscape Physician Compensation Report 2016	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Sample size: 19,183 physicians across 26 specialties met the screening criteria. Recruitment period: November 17, 2015 – February 9, 2016	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10
count	28	27	27	27	27	28	27	27	27	27	1
unique	28	25	23	13	25	16	14	21	22	11	1
top	http://medscape.com/features/slideshow/compens...	$222,000	36%	44%	$56,000	53%	44%	68%	48%	1%	(Labels below are correct only when Annual Inc...
freq	1	2	2	4	2	3	4	3	3	10	1

Activity 2: Visual analysis;

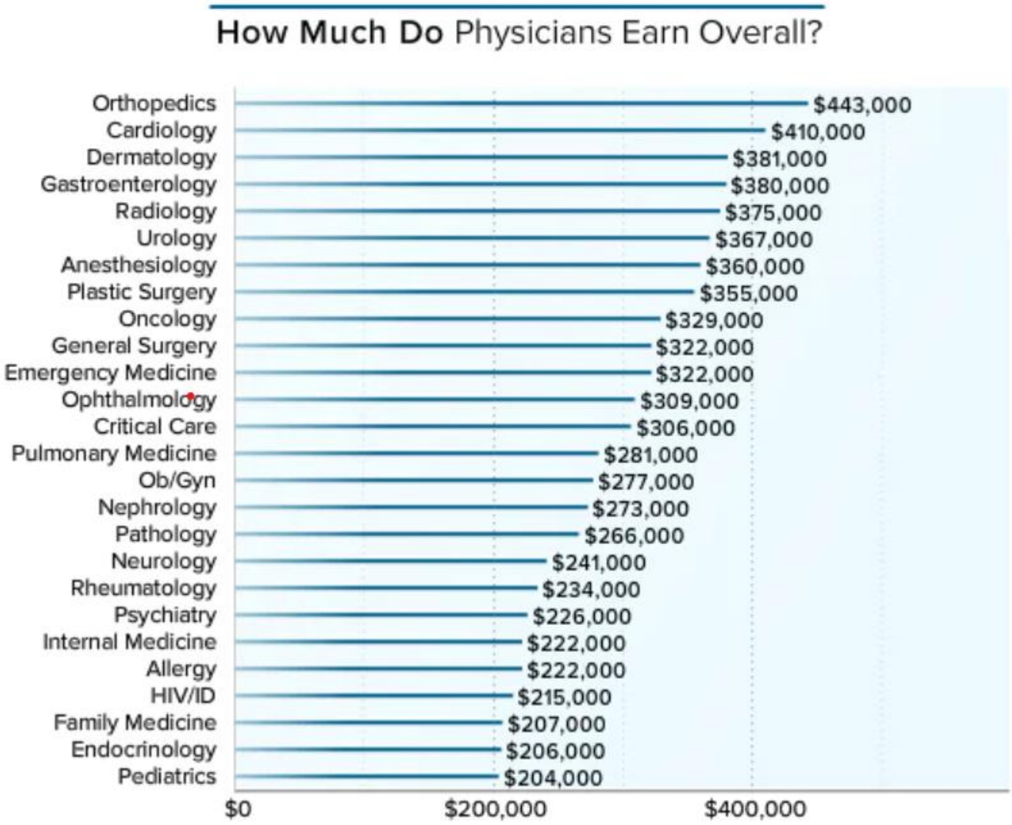
Visual analysis is the process of using visual representations, such as charts, plots, and graphs, to explore and understand data. It is a way to quickly identify patterns, trends, and outliers in the data,
which can help to gain insights and make informed decisions.



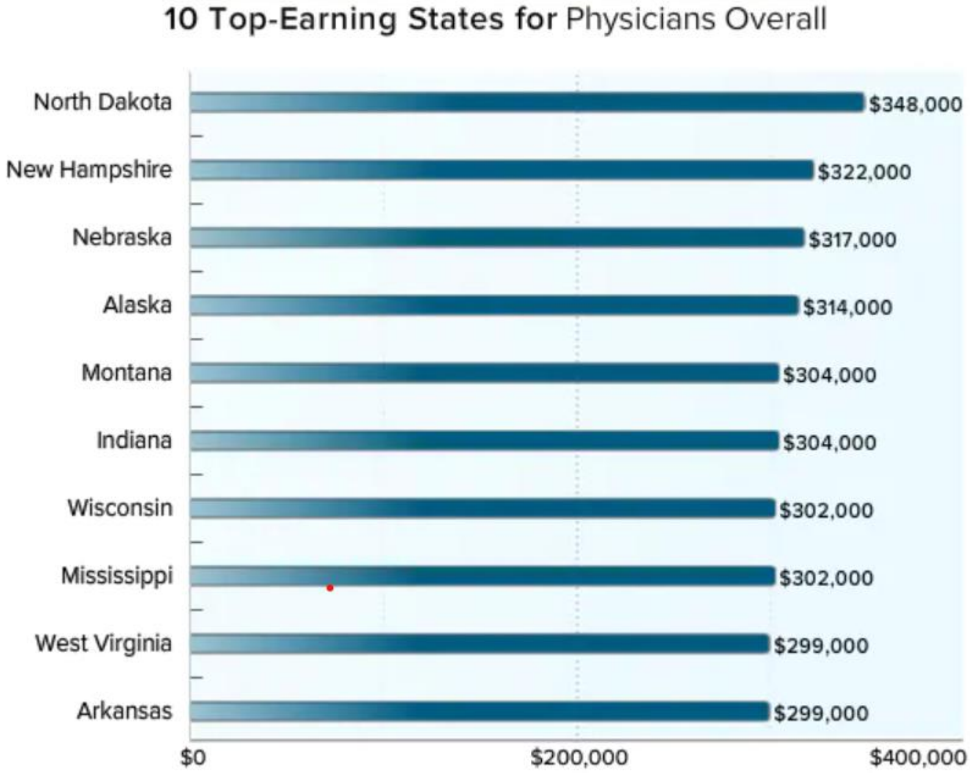
```
from matplotlib import pyplot as plt
import seaborn as sns
_df_0.groupby('Medscape Physician Compensation Report 2016').size().plot(kind='barh', color=sns.palettes.mpl_palette('Dark2'))
plt.gca().spines[['top', 'right',]].set_visible(False)
```

Medscape Physician Compensation Report 2016
http://medscape.com/features/slideshow/compensation/2016/public/overview
28
1
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

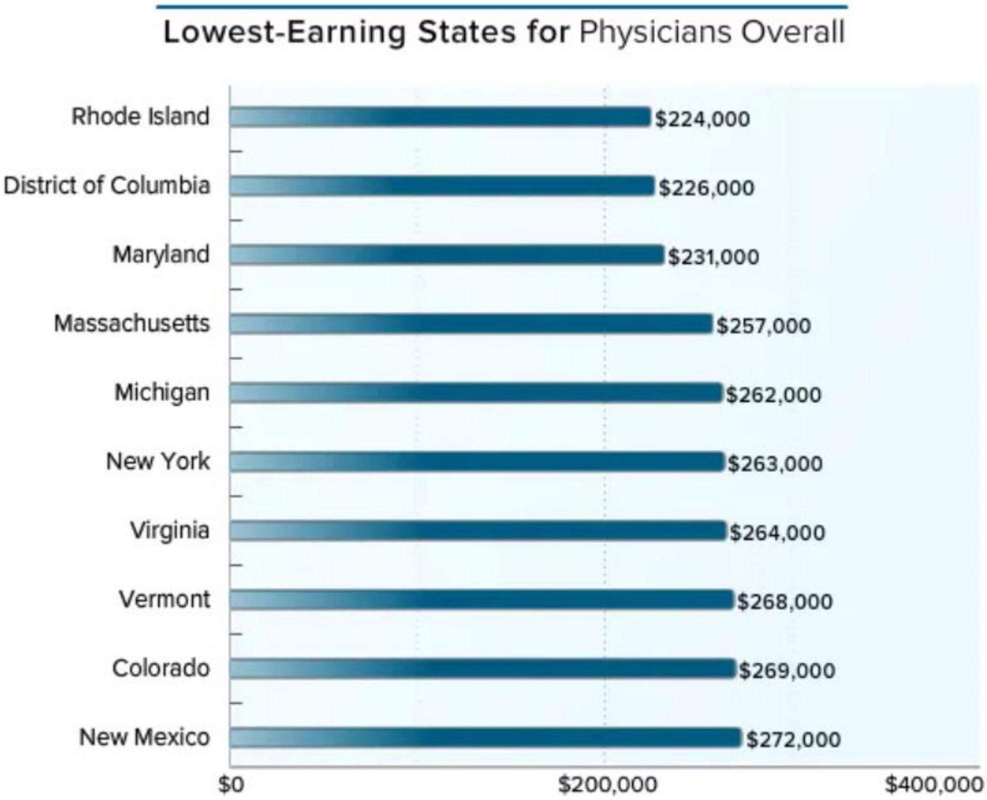
Activity 2.1: How much do doctors earn overall?



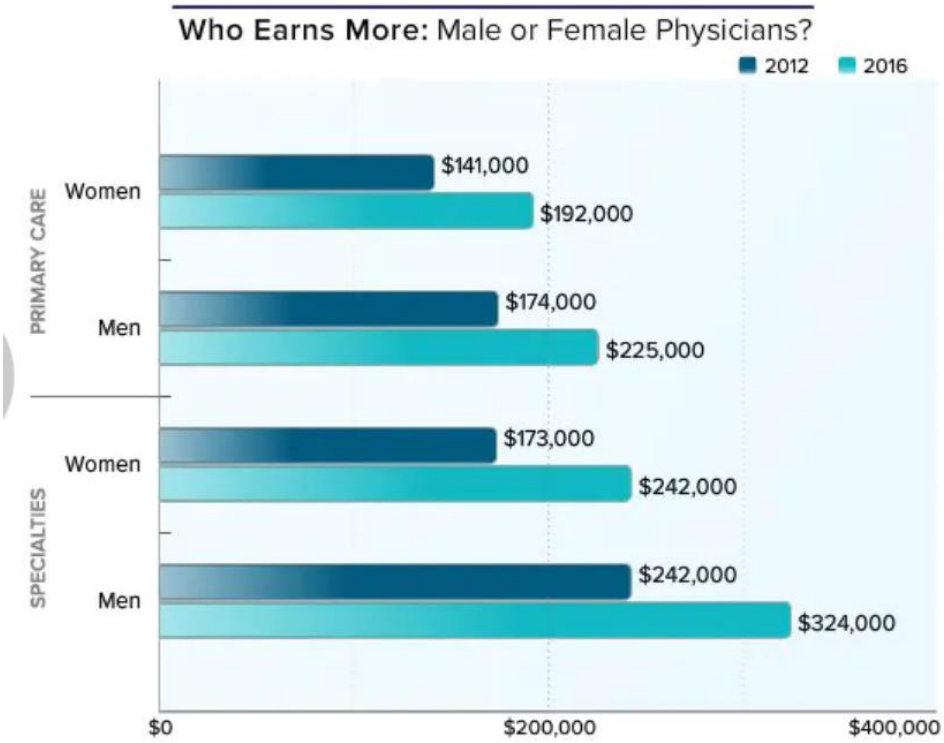
Activity 2.2: Top 10 highest earning doctors;



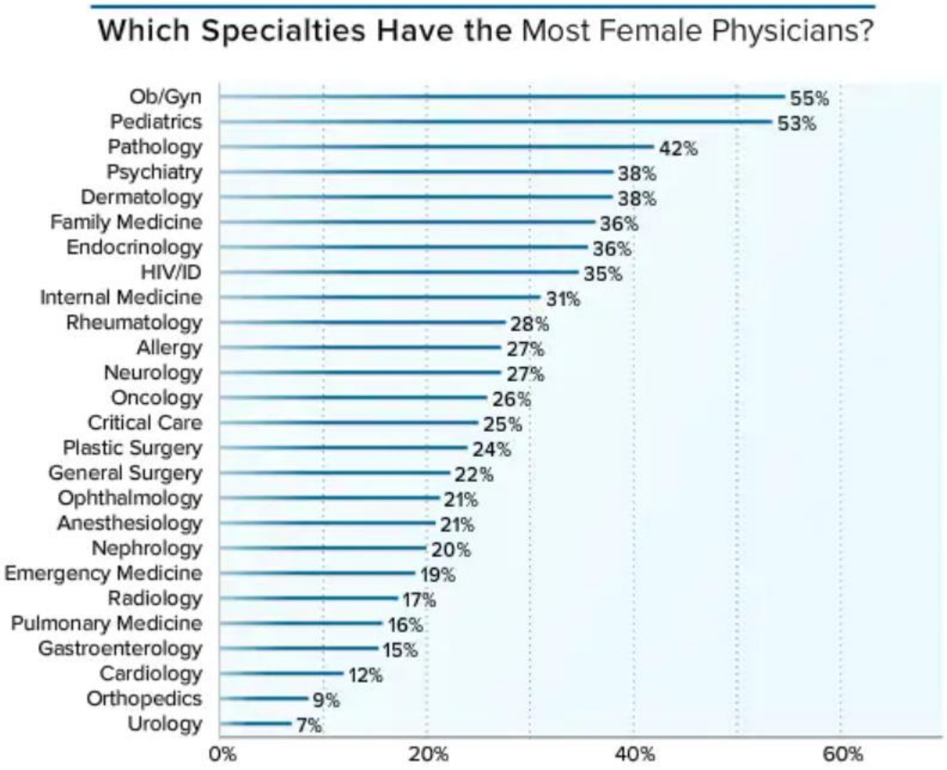
Activity 2.3: Lowest earning doctors;



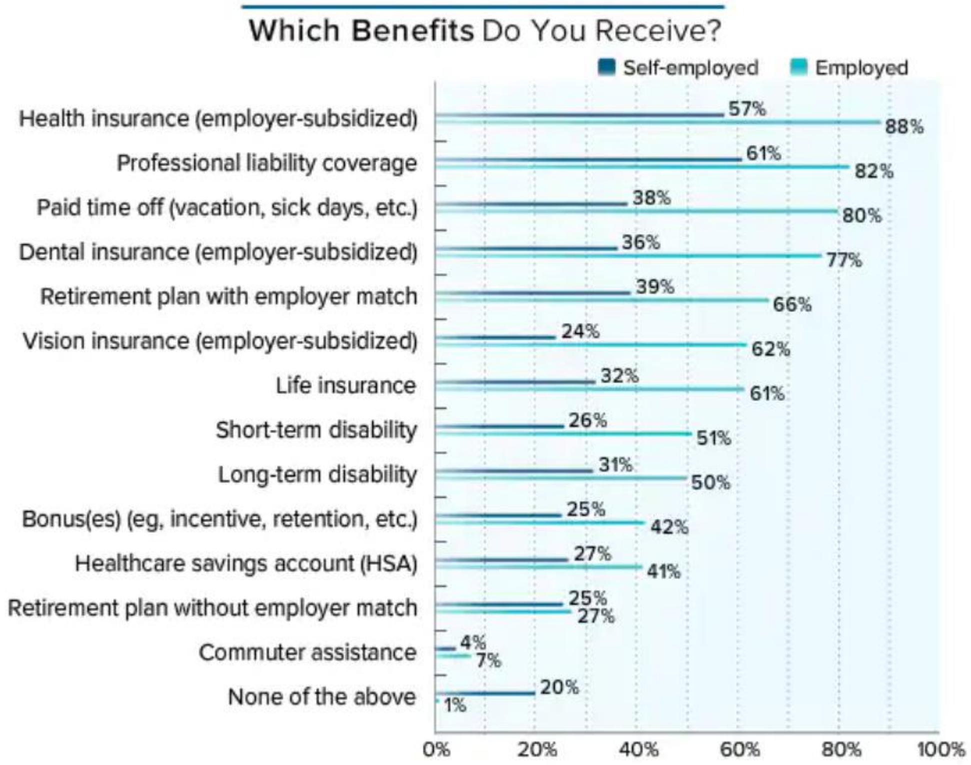
Activity 2.4: Who earns more?



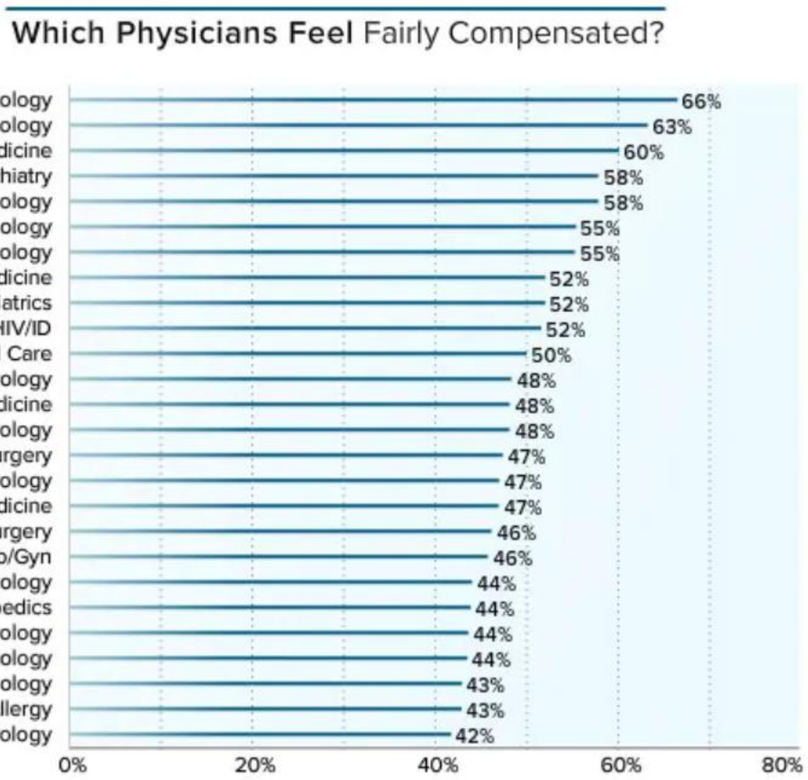
Activity 2.5: Which specialities have the most female doctors?


Which Specialties Have the Most Female Physicians?

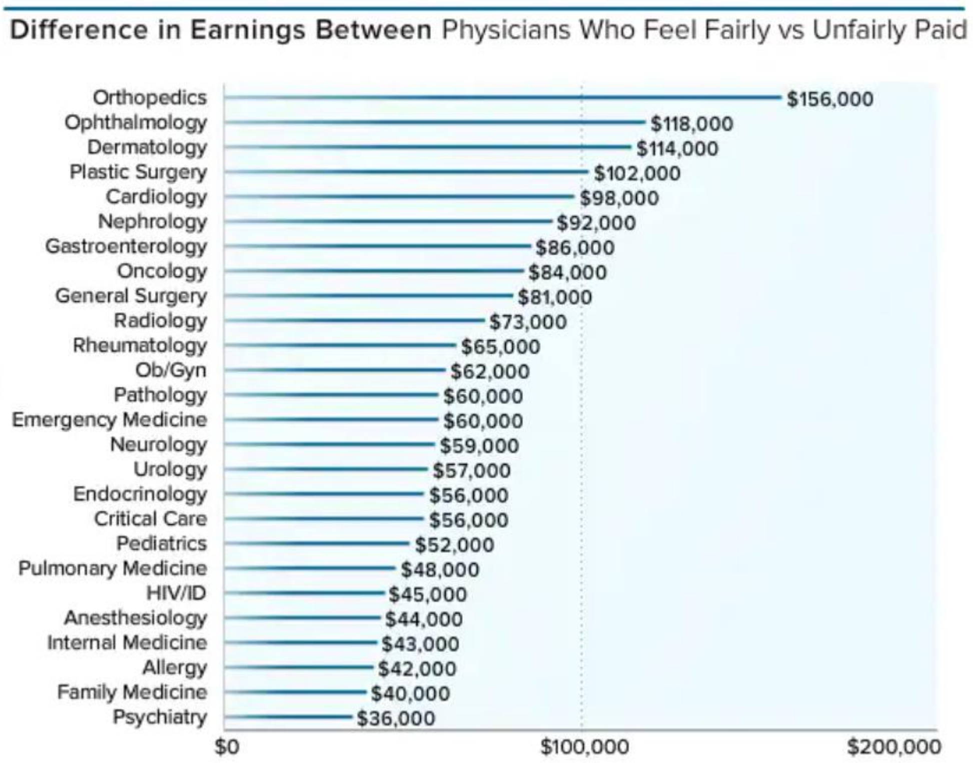
Activity 2.6: What benefits do you get?


Which Benefits Do You Receive?

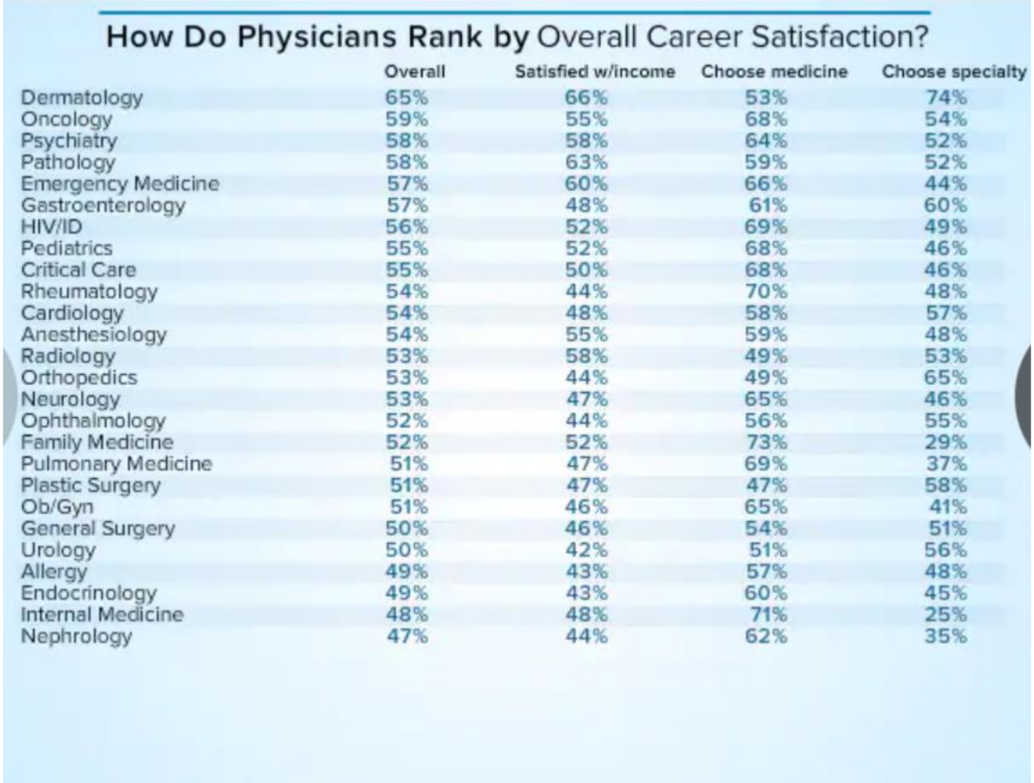
Activity 2.7: Which doctors feel fairly compensated?



Activity 2.8: Difference in earnings between doctors who feel compensated and those who do not.



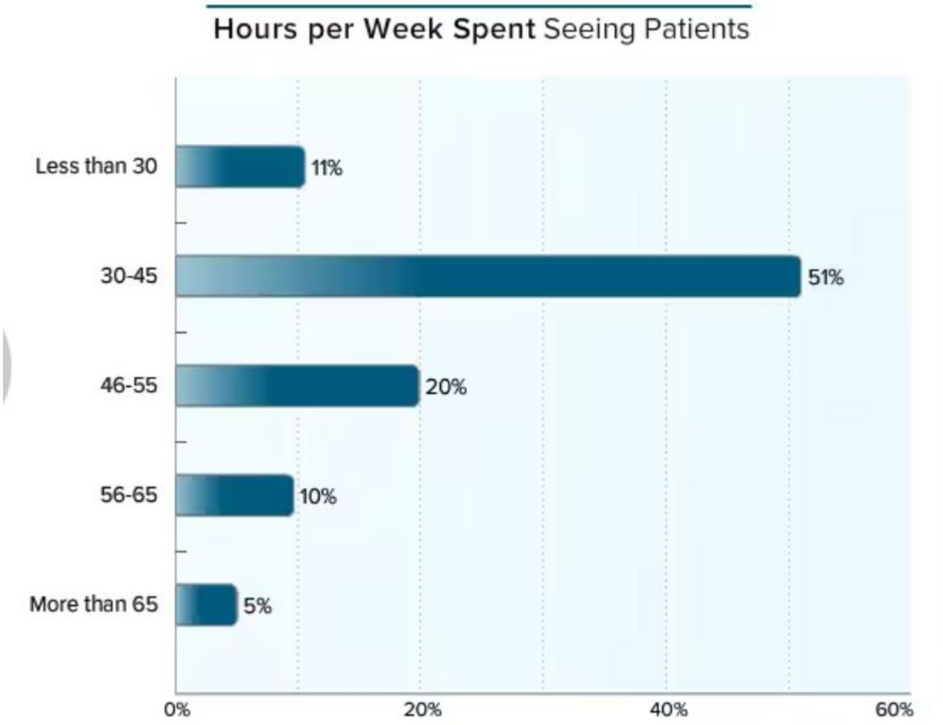
Activity 2.9: Overall Career satisfaction by rank.



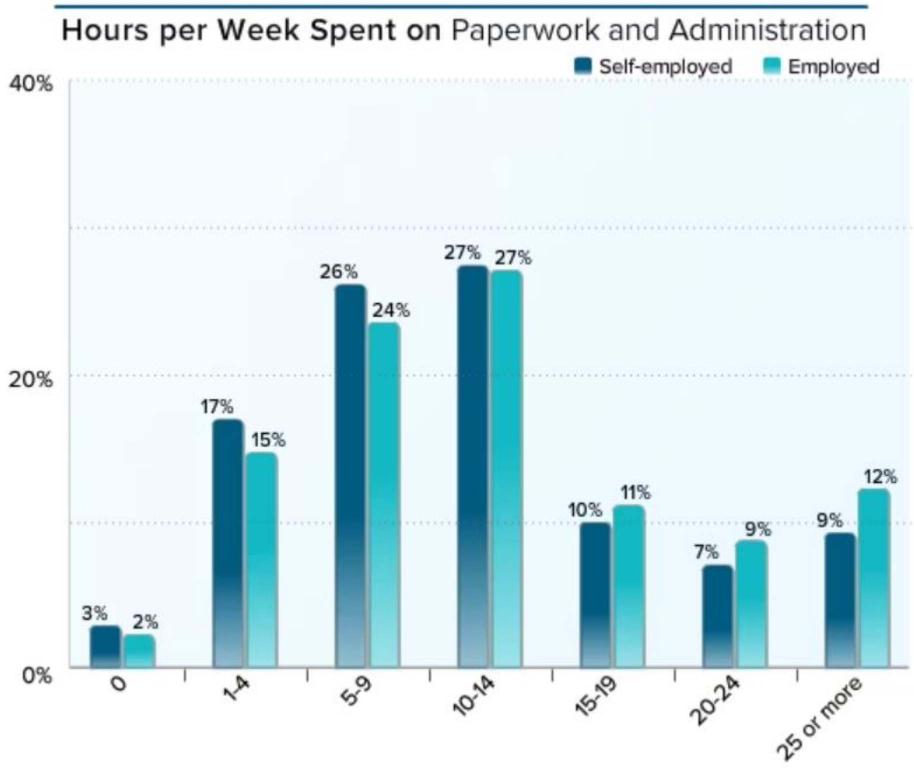
How Do Physicians Rank by Overall Career Satisfaction?

	Overall	Satisfied w/income	Choose medicine	Choose specialty
Dermatology	65%	66%	53%	74%
Oncology	59%	55%	68%	54%
Psychiatry	58%	58%	64%	52%
Pathology	58%	63%	59%	52%
Emergency Medicine	57%	60%	66%	44%
Gastroenterology	57%	48%	61%	60%
HIV/ID	56%	52%	69%	49%
Pediatrics	55%	52%	68%	46%
Critical Care	55%	50%	68%	46%
Rheumatology	54%	44%	70%	48%
Cardiology	54%	48%	58%	57%
Anesthesiology	54%	55%	59%	48%
Radiology	53%	58%	49%	53%
Orthopedics	53%	44%	49%	65%
Neurology	53%	47%	65%	46%
Ophthalmology	52%	44%	56%	55%
Family Medicine	52%	52%	73%	29%
Pulmonary Medicine	51%	47%	69%	37%
Plastic Surgery	51%	47%	47%	58%
Ob/Gyn	51%	46%	65%	41%
General Surgery	50%	46%	54%	51%
Urology	50%	42%	51%	56%
Allergy	49%	43%	57%	48%
Endocrinology	49%	43%	60%	45%
Internal Medicine	48%	48%	71%	25%
Nephrology	47%	44%	62%	35%

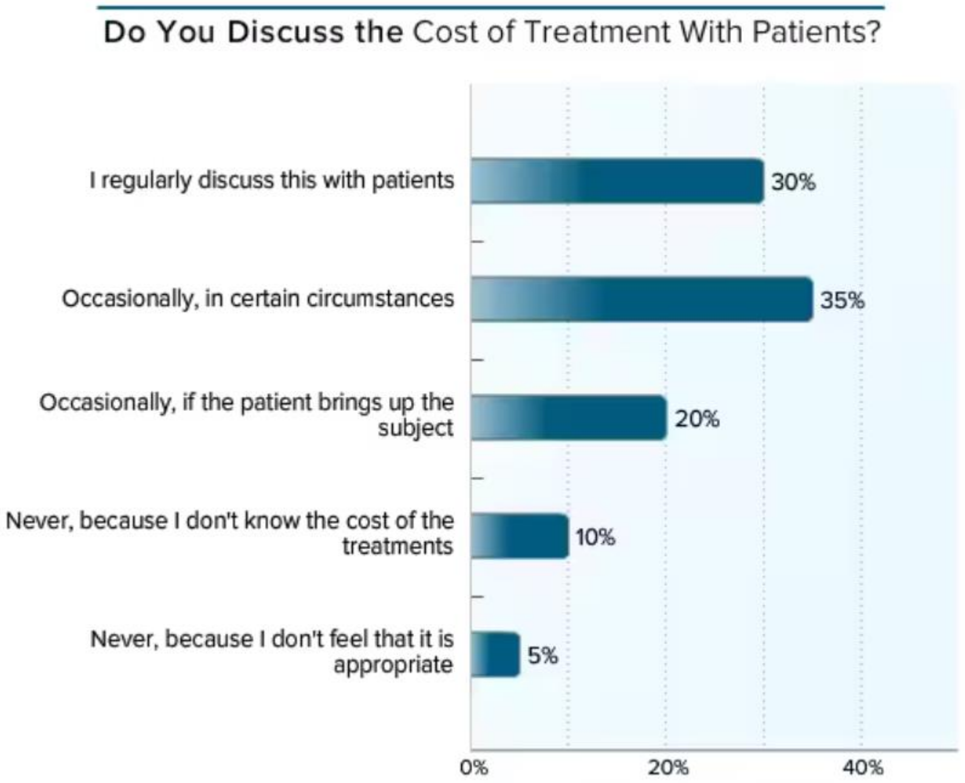
Activity 2.10: Hours spent per week seeing patients.



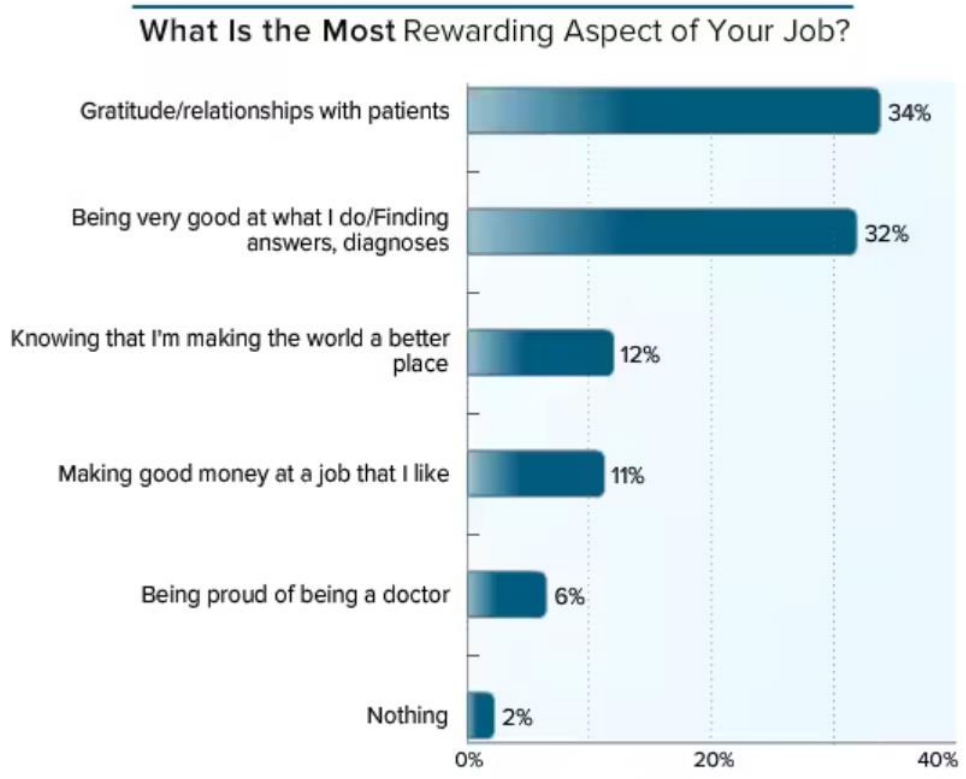
Activity 2.11: Hours spent per week on paper work and administration.



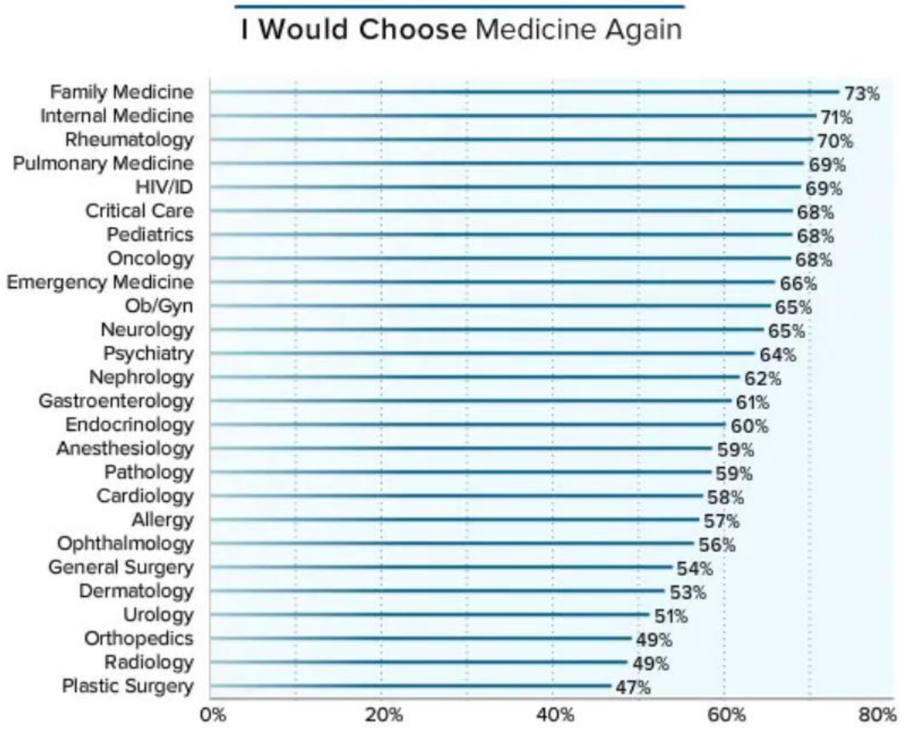
Activity 2.12: Is the cost of treatment discussed with the patients.



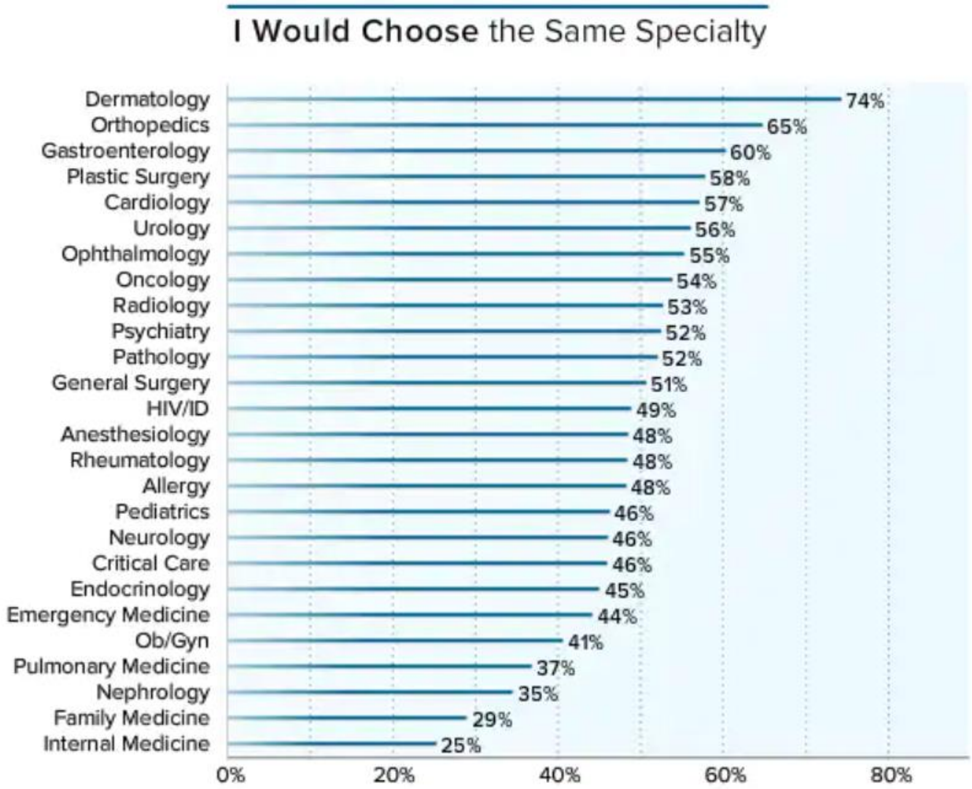
Activity 2.13: Most rewarding aspect of the job?



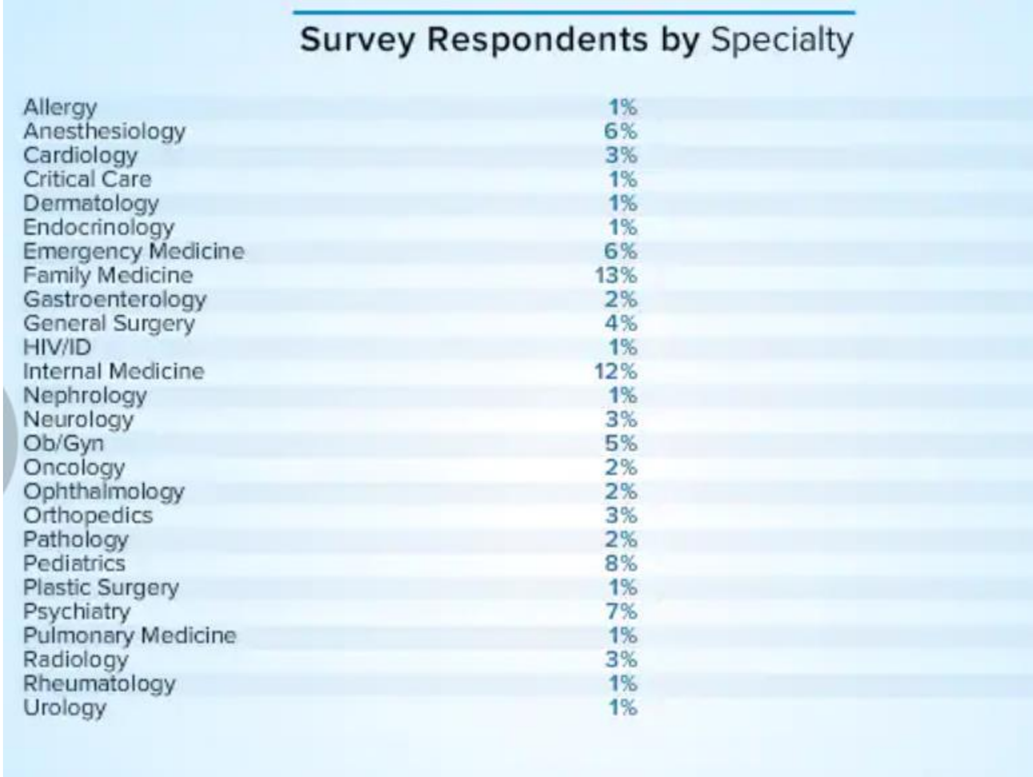
Activity 2.14: Would choose medicine all over again?



Activity 2.15: Would choose the same specialty again?



Activity 2.16: Survey respondents by specialty.



Survey Respondents by Specialty

Allergy	1%
Anesthesiology	6%
Cardiology	3%
Critical Care	1%
Dermatology	1%
Endocrinology	1%
Emergency Medicine	6%
Family Medicine	13%
Gastroenterology	2%
General Surgery	4%
HIV/ID	1%
Internal Medicine	12%
Nephrology	1%
Neurology	3%
Ob/Gyn	5%
Oncology	2%
Ophthalmology	2%
Orthopedics	3%
Pathology	2%
Pediatrics	8%
Plastic Surgery	1%
Psychiatry	7%
Pulmonary Medicine	1%
Radiology	3%
Rheumatology	1%
Urology	1%

Milestone 4: Model Building

Activity 1: Training the model in multiple algorithms

Now our data is cleaned and it’s time to build the model. We can train our data on different algorithms. For this project, we are applying three classification algorithms. The best model is saved based on its performance.

Activity 1.1: Linear Regression

A function named Linear Regression is created and train and test data are passed as the parameters. Inside the function, Linear Regression algorithm is initialised and training data is passed to the model with the .fit() function. Test data is predicted with .predict() function and saved in a new variable. For evaluating the model with R2_score.