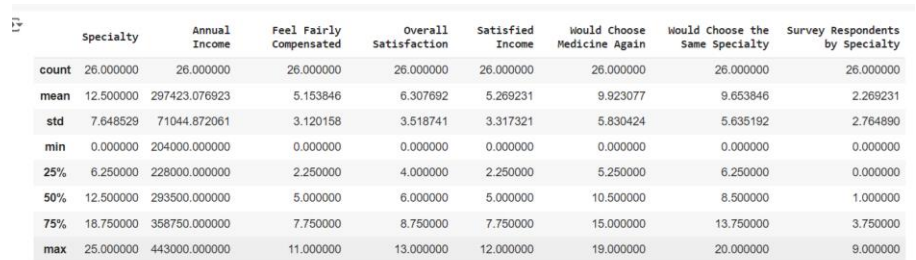


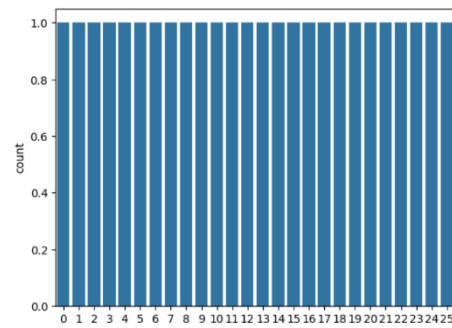
Data Collection and Preprocessing Phase

Date	15 July 2024
Team ID	740054
Project Title	Doctors Annual Salary prediction
Maximum Marks	6 Marks

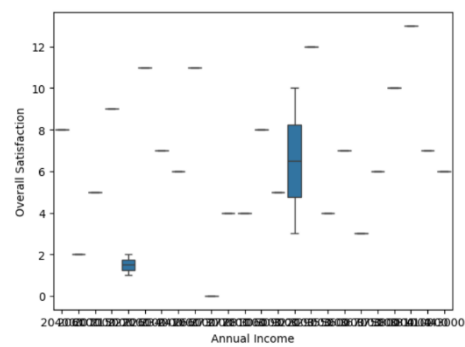
Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

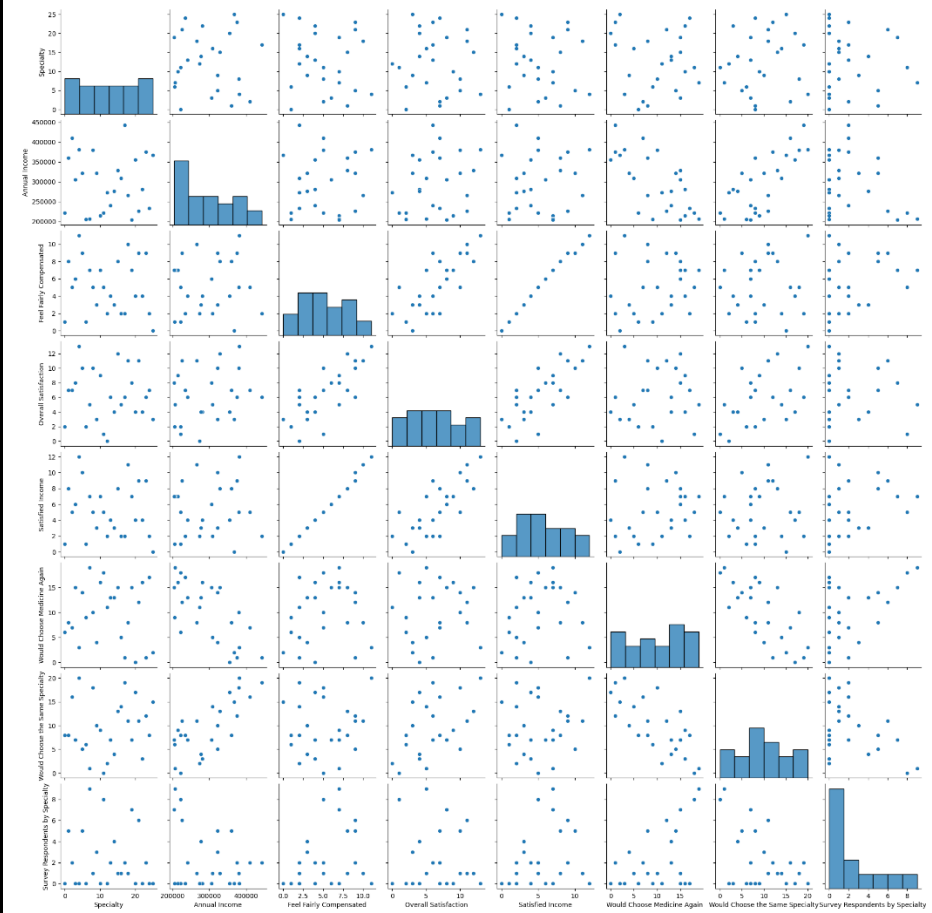
Section	Description
Data Overview	<p><u>Dimension:</u> 28 rows × 8 columns</p> <p><u>Descriptive statistics:</u></p> 
Countplot Analysis	



Boxplot Analysis



PairPlot Analysis



Outliers and Anomalies

-

Data Preprocessing Code Screenshots

Loading Data	<div><pre>df = pd.read_excel('/content/NewDoctorsPay.xlsx') df</pre></div> <table><thead><tr><th></th><th>Specialty</th><th>Annual Income</th><th>Feel Fairly Compensated</th><th>Overall Satisfaction</th><th>Satisfied Income</th><th>Would Choose Medicine Again</th><th>Would Choose the Same Specialty</th><th>Survey Respondents by Specialty</th></tr></thead><tbody><tr><td>0</td><td>Orthopedics</td><td>443000</td><td>0.44</td><td>0.53</td><td>0.44</td><td>0.49</td><td>0.65</td><td>0.03</td></tr><tr><td>1</td><td>Cardiology</td><td>410000</td><td>0.48</td><td>0.54</td><td>0.48</td><td>0.58</td><td>0.57</td><td>0.03</td></tr><tr><td>2</td><td>Dermatology</td><td>381000</td><td>0.66</td><td>0.65</td><td>0.66</td><td>0.53</td><td>0.74</td><td>0.01</td></tr><tr><td>3</td><td>Gastroenterology</td><td>380000</td><td>0.48</td><td>0.57</td><td>0.48</td><td>0.61</td><td>0.60</td><td>0.02</td></tr><tr><td>4</td><td>Radiology</td><td>375000</td><td>0.58</td><td>0.53</td><td>0.58</td><td>0.49</td><td>0.53</td><td>0.03</td></tr><tr><td>5</td><td>Urology</td><td>367000</td><td>0.42</td><td>0.50</td><td>0.42</td><td>0.51</td><td>0.56</td><td>0.01</td></tr><tr><td>6</td><td>Anesthesiology</td><td>360000</td><td>0.55</td><td>0.54</td><td>0.55</td><td>0.59</td><td>0.48</td><td>0.06</td></tr><tr><td>7</td><td>Plastic Surgery</td><td>355000</td><td>0.47</td><td>0.51</td><td>0.47</td><td>0.47</td><td>0.58</td><td>0.01</td></tr><tr><td>8</td><td>Oncology</td><td>329000</td><td>0.55</td><td>0.59</td><td>0.55</td><td>0.68</td><td>0.54</td><td>0.02</td></tr><tr><td>9</td><td>Emergency Medicine</td><td>322000</td><td>0.58</td><td>0.57</td><td>0.60</td><td>0.66</td><td>0.44</td><td>0.06</td></tr><tr><td>10</td><td>General Surgery</td><td>322000</td><td>0.46</td><td>0.50</td><td>0.46</td><td>0.54</td><td>0.51</td><td>0.04</td></tr><tr><td>11</td><td>Ophthalmology</td><td>309000</td><td>0.44</td><td>0.52</td><td>0.44</td><td>0.56</td><td>0.55</td><td>0.02</td></tr><tr><td>12</td><td>Critical Care</td><td>306000</td><td>0.50</td><td>0.55</td><td>0.50</td><td>0.68</td><td>0.46</td><td>0.01</td></tr><tr><td>13</td><td>Pulmonary Medicine</td><td>281000</td><td>0.47</td><td>0.51</td><td>0.47</td><td>0.69</td><td>0.37</td><td>0.01</td></tr></tbody></table>		Specialty	Annual Income	Feel Fairly Compensated	Overall Satisfaction	Satisfied Income	Would Choose Medicine Again	Would Choose the Same Specialty	Survey Respondents by Specialty	0	Orthopedics	443000	0.44	0.53	0.44	0.49	0.65	0.03	1	Cardiology	410000	0.48	0.54	0.48	0.58	0.57	0.03	2	Dermatology	381000	0.66	0.65	0.66	0.53	0.74	0.01	3	Gastroenterology	380000	0.48	0.57	0.48	0.61	0.60	0.02	4	Radiology	375000	0.58	0.53	0.58	0.49	0.53	0.03	5	Urology	367000	0.42	0.50	0.42	0.51	0.56	0.01	6	Anesthesiology	360000	0.55	0.54	0.55	0.59	0.48	0.06	7	Plastic Surgery	355000	0.47	0.51	0.47	0.47	0.58	0.01	8	Oncology	329000	0.55	0.59	0.55	0.68	0.54	0.02	9	Emergency Medicine	322000	0.58	0.57	0.60	0.66	0.44	0.06	10	General Surgery	322000	0.46	0.50	0.46	0.54	0.51	0.04	11	Ophthalmology	309000	0.44	0.52	0.44	0.56	0.55	0.02	12	Critical Care	306000	0.50	0.55	0.50	0.68	0.46	0.01	13	Pulmonary Medicine	281000	0.47	0.51	0.47	0.69	0.37	0.01
	Specialty	Annual Income	Feel Fairly Compensated	Overall Satisfaction	Satisfied Income	Would Choose Medicine Again	Would Choose the Same Specialty	Survey Respondents by Specialty																																																																																																																																
0	Orthopedics	443000	0.44	0.53	0.44	0.49	0.65	0.03																																																																																																																																
1	Cardiology	410000	0.48	0.54	0.48	0.58	0.57	0.03																																																																																																																																
2	Dermatology	381000	0.66	0.65	0.66	0.53	0.74	0.01																																																																																																																																
3	Gastroenterology	380000	0.48	0.57	0.48	0.61	0.60	0.02																																																																																																																																
4	Radiology	375000	0.58	0.53	0.58	0.49	0.53	0.03																																																																																																																																
5	Urology	367000	0.42	0.50	0.42	0.51	0.56	0.01																																																																																																																																
6	Anesthesiology	360000	0.55	0.54	0.55	0.59	0.48	0.06																																																																																																																																
7	Plastic Surgery	355000	0.47	0.51	0.47	0.47	0.58	0.01																																																																																																																																
8	Oncology	329000	0.55	0.59	0.55	0.68	0.54	0.02																																																																																																																																
9	Emergency Medicine	322000	0.58	0.57	0.60	0.66	0.44	0.06																																																																																																																																
10	General Surgery	322000	0.46	0.50	0.46	0.54	0.51	0.04																																																																																																																																
11	Ophthalmology	309000	0.44	0.52	0.44	0.56	0.55	0.02																																																																																																																																
12	Critical Care	306000	0.50	0.55	0.50	0.68	0.46	0.01																																																																																																																																
13	Pulmonary Medicine	281000	0.47	0.51	0.47	0.69	0.37	0.01																																																																																																																																
Data Transformation	<pre>x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42) x_train = x_train.replace('Any', '', regex=True).astype('float') / 100 x_test = x_test.replace('Any', '', regex=True).astype('float') / 100 y_train = y_train.replace('Any', '', regex=True).astype('float') / 100 y_test = y_test.replace('Any', '', regex=True).astype('float') / 100 imputer_x = SimpleImputer(strategy='mean') x_train = pd.DataFrame(imputer_x.fit_transform(x_train)) x_test = pd.DataFrame(imputer_x.transform(x_test)) imputer_y = SimpleImputer(strategy='mean') y_train = imputer_y.fit_transform(y_train.values.reshape(-1, 1)) y_test = imputer_y.transform(y_test.values.reshape(-1, 1)) reg = LinearRegression() reg.fit(x_train, y_train)</pre>																																																																																																																																							
Feature Engineering	Attached the codes in final submission.																																																																																																																																							
Save Processed Data	-																																																																																																																																							