ANALYSIS ON WORKPLACE CONDITIONS DURING PANDEMI

Submitted by:
Rahul Barpat
PGDM,
IPE,
Shamirpet.

Objective

Main Objective:

To measure the satisfaction level of employees with the changing workplace.

Sub objectives:

1. To measure the general satisfaction of employees with the current workplace
2. To measure the level of satisfaction/ dissatisfaction of employees with regards to benefits and compensation
3. To assess the satisfaction level of employees with regards to the performance management process at the workplace
4. To evaluate the satisfaction level of employees in relation to work environment of the company
5. To investigate the experience of employees during pandemic

Research Methodology

The population for the study was drawn from various employees working in the public and private sectors. The concept of sampling comes from selecting a specific number of participants from a larger population. The data collection method allows for the selection of non-random willing participants who can assist with the survey process. The research was explored using convenience sampling, selecting to invite participants who were available and eager to participate in the study.

Sampling Frame:

1. Convenience Sampling is a type of non-probability sampling that involves the sample being drawn from that part of the population that is close to hand.
2. Judgemental Approach- friends who have work experience during pandemic period.
3. Simple Random Sampling- Randomly selected samples

- Private Sectors Respondents.

The time period for the survey was considered as employees working in pandemic i.e., from April 2020 to till date. The research ensured the confidentiality of the research participants identities. Later, the data obtained from the survey were exported into EXCEL and examined further.

Instrument: The job satisfaction survey included factors such as

- Satisfaction with pay
- Fringe benefits
- Contingent rewards
- Supervision
- Nature of work
- Communication
- Operational procedures, etc.

Research Hypothesis

To determine the association between the variables, statistical tools like contingency tables have been used. Following hypothesis have been formulated to determine the significance. If p vale is less than or equal to 0.05, reject H_0.

1. H_0: There is no association between gender and satisfaction level during pandemic [In 22]

   H_1: There is an association between gender and satisfaction level during pandemic [In 22]

2. H_0: There is no association between age and satisfaction level of the employees during pandemic [In 24]

   H_1: There is an association between age and satisfaction level of the employees during pandemic [In 24]

3. H_0: There is no association between salary and satisfaction level of the employees during pandemic [In 25]

   H_1: There is an association between salary and satisfaction level of the employees during pandemic [In 25]

4. H_0: There is no association between gender and benefits provided by the employer [In 26]

   H_1: There is an association between gender and benefits provided by the employer [In 26]

5. H_0: There is no association between age and benefits provided by the employer [In 27]

   H_1: There is an association between age and benefits provided by the employer [In 27]

6. H_0: There is no association between salary and benefits provided by the employer [In 28]

   H_1: There is an association between salary and benefits provided by the employer [In 28]

7. H_0: There is no association between gender and personal performance and career goals of the employees [In 29]

H_1: There is an association between gender and personal performance and career goals of the employees [In 29]

8. H_0: There is no association between age and personal performance and career goals of the employees [In 30]

   H_1: There is an association between age and personal performance and career goals of the employees [In 30]

9. H_0: There is no association between salary and personal performance and career goals of the employees [In 31]

   H_1: There is an association between salary and personal performance and career goals of the employees [In 31]

10. H_0: There is no association between gender and work opportunities for the employees [In 32]

    H_1: There is an association between gender and work opportunities for the employees [In 32]

11. H_0: There is no association between age and work opportunities for the employees [In 33]

    H_1: There is an association between age and work opportunities for the employees [In 33]

12. H_0: There is no association between salary and work opportunities for the employees [In 34]

    H_1: There is an association between salary and work opportunities for the employees [In 34]

13. H_0: There is no association between gender and work culture and discrimination present in the organization [In 35]

    H_1: There is an association between gender and work culture and discrimination present in the organization [In 35]

Data Sources

Structured Data:

Structured data is the data which conforms to a data model, has a well define structure, follows a consistent order, and can be easily accessed and used by a person or a computer program.

Structured data is usually stored in well-defined schemas such as Databases. It is generally tabular with column and rows that clearly define its attributes.

In this project work we have used Structured Questionnaire to collect data. Below is the questionnaire from which data has been collected.

Structured Questionnaire:

A Study to Measure the Satisfaction Levels of Employees with Changing Workplace

Name of the Respondent	
Gender	
Age	
Qualification	
Total Experience (in months)	
Organization Name	
Designation	
Experience with Current Organization (in months)	
Salary	

1. Please express your level of agreement/ disagreement for the following statements:
   (1 – Strongly Agree, 2 – Agree, 3 – Neutral, 4 – Disagree; 5 – Strongly Disagree)

	1	2	3	4	5
If I do a good job I will be rewarded					
I get what I need to do my job well					
The conditions I work in are good					
I feel stress in my job					
I am interested in my job					
Management makes wise decisions					
I am proud to work for this company					

2. Please express your level of satisfaction/ dissatisfaction with the employee benefits & compensation during pandemic:
   (1 – Poor, 2 – Fair, 3 – Good, 4 – Very Good; 5 – Excellent)

	1	2	3	4	5
Health Care					
Vacation Time					
Training Availability					
Performance Reviews					
Day Care					
Parking					
Retirement Plan					
Salary/Wages					

3. How often do you undergo the following review processes:

   (1 – Weekly, 2 – Monthly, 3 – Quarterly, 4 – Annually; 5 – Never)

	1	2	3	4	5
Personal Performance					
Departmental Performance					
Company Performance					
Career Goals					

4. Please express your level of agreement or disagreement for the following statements:

   (1 – Strongly Disagree, 2 – Disagree, 3 – Neutral, 4 – Agree; 5 – Strongly Agree)

	1	2	3	4	5
Company tries to create an exciting work environment					
I receive enough opportunity to interact with other employees on a formal level					
I receive enough opportunity to interact with other employees on an informal level					
I feel a sense of completion with my job					
I have enough freedom in my position to take independent action when needed					
I have enough freedom in my position to do what is right for the customer					
I have a clear path for career advancement					

5. Overall how satisfied are you with your position at this company during pandemic?
   - □ Extremely Dissatisfied
   - □ Dissatisfied
   - □ Neutral
   - □ Satisfied
   - □ Extremely Satisfied

6. How flexible is the company with respect to your family responsibilities?
   - □ Extremely Inflexible
   - □ Inflexible
   - □ Neutral
   - □ Flexible
   - □ Extremely Flexible

7. Do you take part in your company's flexi-time program?
   - □ Yes
   - □ No

8. Have you ever observed or experienced any of the following forms of discrimination or harassment at this company?
   - □ Racial discrimination
   - □ Sexual orientation discrimination
   - □ Gender discrimination
   - □ Age discrimination
   - □ None of the above

9. My job requirements are clear.
   - □ Extremely Dissatisfied
   - □ Dissatisfied
   - □ Neutral
   - □ Satisfied
   - □ Extremely Satisfied

10. How would you rate your reporting manager in each of the following areas during work from home?

(1 – Poor, 2 – Fair, 3 – Good, 4 – Very Good; 5 – Excellent)

	1	2	3	4	5
Communication					
Planning and organizing					
Directing and coordinating					
Job/Technical knowledge					
Interpersonal relationship					

Personal Interview:

A personal or face-to-face interview employs a standard structured questionnaire (or interview schedule) to ensure that all respondents are asked questions in the same sequences.

It is a two-way conversation initiated by an interviewer to obtain information from a respondent. The questions, the wording, and the sequence define the structure of the interview, and the interview is conducted face-to-face.

Here are the questions which are asked in the personal interview.

1) Are you satisfied with the job, work culture, and management of the company? Please share your best, good and worst experiences.

2) Does your company provide good benefits?

3) How often does your company undergo the review process at all levels?

4) How is your interaction with other employees?

5) How satisfied are you with your job position in the company during the pandemic?

6) How is your work-life balance during the pandemic?

7) Do you take part in the company's flexi-time program?

8) Is there any discrimination in your company?

9) Is the Job requirement clear?

10) How helpful was your reporting manager during work from home?

Observations and Interpretation of PI:

1. Are you satisfied with the job, work culture, and management of the company? Please share your best, good and worst experiences.

A. Yes, I am fully satisfied with the job, work culture and management of the company. I got great peers, and the organization believes in continuous learning so I could learn more new things during my job.

2. Does your company provide good benefits?

A. Yes, company do provide good benefits like it provides great bonus, and there are few allowances that are for buying electronics and many more.

3. How often does your company undergo the review process at all levels?

A. Company undergoes the review process once in 6 months at all levels.

4. How is your interaction with other employees?

A. The company has got high professionals who are welcoming and more communicative. They guided me with my tasks and helped me build skill sets.

5. How satisfied are you with your job position in the company during the pandemic?

A. I am well satisfied with my job position in the company during the pandemic because there is no change in the pay, I get and the work I do. I just could not work form office is only the thing. Everything else was good.

6. How is your work-life balance during the pandemic?

A. I could not give much time for myself in the pandemic due to random office hours. I could not maintain work-life balance.

7. Do you take part in the company's flexi-time program?

A. Yes, during pandemic we had this kind of program so that we work as per our situation. I really liked this process.

8. Is there any discrimination in your company?

A. No, there is no discrimination in the company. The company provides fair opportunities to everyone irrespective of gender and cast. It even supports LGBTQIA to grow and improve their career.

9. Is the Job requirement clear?

A. Yes, the job requirement is clear, and it is discussed clearly during the recruitment process.

10. How helpful was your reporting manager during work from home?

A. My reporting manager was flexible with the tasks. He kept motivating me and gave me emotional support to complete my task. He guided me ways to do my tasks efficiently and do it quickly without taking much pressure.

Focus Group Discussion:

A focus group discussion involves gathering people from similar backgrounds or experiences together to discuss a specific topic of interest. It is a form of qualitative research where questions are asked about their perceptions attitudes, beliefs, opinion, or ideas. In focus group discussion participants are free to talk with other group members; unlike other research methods it encourages discussions with other participants.

Questions:

1) Can you please tell your overall experience while working during covid time? Was it good or bad?

2) Did you have work life balance during covid time?

Observations:

- Most of them have worked from home during covid 19 where they had a good experience while other few of them have a bad experience.

- People with good experience mentioned they managed things by themselves, didn’t have to skip things from their routine. They had time to work out. It seemed fun for them to sit in comfy pajamas all day and work on the comfort of their bed.
- They save a lot of money which they use to waste like spending on food, clothes, petrol etc. And whenever they don’t have much work, they can do other households’ work.
- However Gradually, they realized that dressing up formally, commuting to work, and working inside the office somehow works great to add to their professionalism.
- On the other hand, there are people who like to leave the work and tension at office and be free when they are home.
- They mentioned it’s kind of hectic as there were no specific timings, people can ping any time and expect work sometimes.
- Some people may not be able to concentrate on work while working from home. And Home atmosphere may not support for people with kids or with more people in home.
- However, with our survey conducted, its observed that approximately 70% of people would like to continue working from home as they get more productive being in their comfort zone.
- Few of them had a great work life balance during Covid.
- They have mentioned that they got to have some family time and able to enjoy all the things that their family members used to do while they were away.
- Just the small things like laughing together when something funny happens, the gossips and all.
- Although initially working from home was a positive factor that will promote work–life balance.
- However, also negative trends appeared, as they were only one call or message away from their employer, and it was therefore expected that they would work outside working hours and would also be available outside working hours. Uncertainty and spending time with family often caused more stress.
- There could be many distractions and if you don't have a proper schedule, you may end ruining the entire work and family time.
- But if you can manage it well you may have the best work life balance.
- With our survey conducted, it was observed that majority of the employees had not much great work life balance during covid.

Statistical tools

Frequency Distribution:

Frequency distributions tell us how frequencies are distributed over the values. That is how many values lie between different intervals.

They give us an idea about the range where most of the values fall and the ranges where values are scarce.

A frequency distribution is an overview of all values of some variable and the number of times they occur.

Contingency Table:

Estimations like mean, median, standard deviation, and variance are very much useful in case of the univariate data analysis.

But in the case of bivariate analysis (comparing two variables) correlation comes into play.

Contingency Table is one of the techniques for exploring two or even more variables. It is basically a tally of counts between two or more categorical variables.

A contingency table, sometimes called a two-way frequency table, is a tabular mechanism with at least two rows and two columns used in statistics to present

categorical data in terms of frequency counts. More precisely, an r×c contingency table shows the observed frequency of two variables, the observed frequencies of

which are arranged into r rows and c columns. The intersection of a row and a column of a contingency table is called a cell.

Chi Square Test:

The Chi-square test is intended to test how likely it is that an observed distribution is due to chance. It is also called a "goodness of fit" statistic,

because it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.

A Chi-square test is designed to analyze categorical data. That means that the data has been counted and divided into categories. It will not work with parametric or continuous data (such as height in inches). For example, if you want to test whether attending class influences how students perform on an exam, using test

scores (from 0-100) as data would not be appropriate for a Chi-square test. However, arranging students into the categories "Pass" and "Fail" would.

Additionally, the data in a Chi-square grid should not be in the form of percentages, or anything other than frequency (count) data.

Null Hypothesis:

The null hypothesis states that there is no statistical significance exists between sets of data which implies that the population parameter will be equal to a hypothesized value.

The null hypothesis assumes that any kind of difference between the chosen characteristics that you see in a set of data is due to chance.

For example, if the expected earnings for the gambling game are truly equal to zero, then any difference between the average earnings in the data and zero is due to chance.

Categorical Plots:

Plots are basically used for visualizing the relationship between variables. Those variables can be either be completely numerical or a category like a group,

class or division. This article deals with categorical variables and how they can be visualized using the Seaborn library provided by Python.

Seaborn besides being a statistical plotting library also provides some default datasets. We will be using one such default dataset called 'tips'.

The 'tips' dataset contains information about people who probably had food at a restaurant and whether they left a tip for the waiters, their gender, whether they smoke and so on.

Numerical Plots:

Numerical data represent values that can be measured and put into a logical order. Examples of numerical data are height, weight, age, number of movies watched,

IQ, etc. To graph numerical data, one uses dot plots, stem and leaf graphs, histograms, box plots, ogive graphs, and scatter plots.

Descriptive Analysis:

Python Descriptive Statistics process describes the basic features of data in a study. It delivers summaries on the sample and the measures and does not use the data to learn about the population it represents.

Under descriptive statistics, fall two sets of properties- central tendency and dispersion. Python Central tendency characterizes one central value for the entire distribution.

Measures under this include mean, median, and mode. Python Dispersion is the term for a practice that characterizes how apart the members of the distribution are from the center and from each other. Variance/Standard Deviation is one such measure of variability.

describe() is used for Exploratory Data Analysis

- Function for Exploratory Analysis or Descriptive/ Summary Statistics
- describe(include='all') - Summary statistics for both numeric/ categorical data items/ features
- describe() - Default EDA - Calculates summary statistics for numerical features
  count, mean, standard deviation, minimum, maximum, 25%, 50%, 75% quantiles
- describe(include=['object']) - Summary statistics for categorical columns (object data type)
  count, unique, top, frequency

- data.describe(include='all')   # EDA for numeric & categorical features too
- data.describe(include=['int64'])   # EDA for int64 features
- data.describe(include=['float64']) # EDA for float64 features
- data.describe()   # EDA for numeric features only
- data.describe(include=['object']) # EDA for categorical features

Diagnostic Analysis:

Diagnostic analytics is the process of using data to determine the causes of trends and correlations between variables. It can be viewed as a logical next step after using descriptive analytics to identify trends. Diagnostic analysis can be done manually, using an algorithm, or with statistical software (such as Microsoft Excel).

There are several concepts to understand before diving into diagnostic analytics: hypothesis testing, the difference between correlation and causation, and diagnostic regression analysis.

The Explanation of the root cause behind the outcome is considered under descriptive analytics.

1) Correlation coefficient is denoted by r

- If [r=1], features have a perfect positive correlation i.e., if one variable moves a given amount, the second moves proportionally in the same direction
- If [r=0], no relationship exists between the features. If one variable moves, you can make no predictions about the movement of the other variable; they are uncorrelated
- If [r=–1], the variables are perfectly negatively correlated (or inversely correlated) & move in opposition to each other. If one variable increases, the other variable decreases proportionally

2) Measuring correlation graphically using scatter_matrix() & heatmap() functions

- Correlation = 0 - No relationship exists between the features
- Correlation = 1 - Perfect positive relationship
- Correlation = -1 - Perfect negative relationship

3) Correlation in Python:

1. corr() - Karl Pearson's Coefficient of Correlation
    - Selects only numeric features for further analysis
    - num_feature - Set of numeric features - int64 or float64 - 12 features (11 are int64 & 1 is float64)
2. scatter() - Graphically plot 2 variables
    - Plot independent variable on x-axis & dependent variable on y axis
    - Limitation - Only 2 variables can be plotted
    - scatter_matrix() - diagonal - kde = kernel density estimation
    - hist = histogram
3. pairplot() - Measures association between multiple features
4. heatmap() - Measures association between multiple features

Simple Linear Regression Model:

It is used to estimate the relationship between two quantitative variables. You can use simple linear regression when you want to know:

How strong the relationship is between two variables (e.g., the relationship between rainfall and soil erosion).

The value of the dependent variable at a certain value of the independent variable (e.g., the amount of soil erosion at a certain level of rainfall).

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Multi Linear Regression Model:

Multiple regression is like linear regression, but with more than one independent value, meaning that we try to predict a value based on two or more variables.

Multiple Linear Regression attempts to model the relationship between two or more features and a response by fitting a linear equation to observed data.

The steps to perform multiple linear Regression are almost like that of simple linear Regression. The Difference Lies in the evaluation. We can use it to find out which factor has the highest impact on the predicted output and how different variables relate to each other.

Analysis and Interpretation

In [1]:
```
#pip install pingouin
#conda install -c conda-forge pingouin
#conda install -c conda-forge pingouin
#pip install --upgrade pingouin
```

In [2]:
```
import numpy as np
import pandas as pd
import ast
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
from plotly import graph_objects as go
import pingouin as pg
```

In [3]:
```
df = pd.read_excel(r"C:\Users\Home\Downloads\Python for Datascience and ML\Final\Team3_Dataset.xlsx")
df
```

Out[3]:

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	...	v37	v38	v39	v40	v41	v42	v43	v44	v45	v46
0	1	Shravya	Female	25	B.Tech	24	Amazon	ROC transportation Specialist	12	500000	...	4	5	1	5	4.0	5	5	5	4	5
1	2	Venkata Ananda Sai Tunuguntala	Male	23	B.Tech	24	WiSig Networks	Product development Engineer	7	650000	...	4	5	2	5	5.0	5	5	5	5	5
2	3	REMELLA SATYA RAMANA KUMAR	Male	24	B.Tech	20	Infosys	Senior system Engineer	20	28000	...	4	4	1	5	4.0	5	4	4	3	4
3	4	R Sai Surya Siva Prasad	Male	23	Masters	12	Accenture	Software Engineer	3	800000	...	3	3	2	3	3.0	2	4	3	2	2
4	5	AKSHAY CH	Male	25	B.Tech	16	Cigniti tech2logies	Associate engineer	3	400000	...	4	4	1	5	4.0	4	4	4	3	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98	99	Leeladhar Reddy Madireddy	Male	25	B.Tech	10	TCS	Asst. System Engineer	10	600000	...	4	3	1	5	4.0	4	3	4	4	3
99	100	Akhil Maddu	Male	25	B.Tech	43	Accenture	Application Development Senior Analyst	8	800000	...	4	3	2	5	5.0	4	3	4	3	2
100	101	Mallika	Female	23	B pharm	24	Medico Health Care	Ar Caller	22	250000	...	3	1	1	4	3.0	3	1	1	3	3
101	102	Sai simha	Male	29	PGDM	4	Tata aig	Associate Management Trainee	4	35000	...	3	5	2	5	5.0	5	5	5	5	5
102	103	Dheeraj PK	Male	26	Post graduate	60	Mediyea	Executive	48	2800000	...	5	5	1	5	4.0	4	4	4	4	4

103 rows × 46 columns

In [4]:
```
df_General_Satisfaction      = df[['v11', 'v12', 'v13', 'v14', 'v15', 'v16', 'v17']]
df_Benefits_and_Compensation = df[['v18', 'v19', 'v20', 'v21', 'v22', 'v23', 'v24', 'v25']]
df_Performance_Management    = df[['v26', 'v27', 'v28', 'v29']]
df_Job_Satisfaction          = df[['v30', 'v31', 'v32', 'v33', 'v34', 'v35', 'v36']]
df_Satisfaction_Pandemic     = df[['v37', 'v38', 'v39', 'v40', 'v41']]
df_Managers_Support          = df[['v42', 'v43', 'v44', 'v45', 'v46']]
```

In [5]:
```
print("Employee Satisfaction - General Facilities          :",pg.cronbach_alpha(data=df_General_Satisfaction))
print("Employee Satisfaction - Benefits & Compensation     :",pg.cronbach_alpha(data=df_Benefits_and_Compensation))
print("Employee Satisfaction - Performance Management      :",pg.cronbach_alpha(data=df_Performance_Management))
print("Employee Satisfaction - Job Role & Responsibilities :",pg.cronbach_alpha(data=df_Job_Satisfaction))
print("Employee Satisfaction - During Pandemic             :",pg.cronbach_alpha(data=df_Satisfaction_Pandemic))
print("Employee Satisfaction - Reporting Manager's Support :",pg.cronbach_alpha(data=df_Managers_Support))
```

```
Employee Satisfaction - General Facilities          : (0.8843712419365434, array([0.847, 0.916]))
Employee Satisfaction - Benefits & Compensation     : (0.9074262646807782, array([0.878, 0.932]))
Employee Satisfaction - Performance Management      : (0.8426620908779916, array([0.786, 0.887]))
Employee Satisfaction - Job Role & Responsibilities : (0.9289121315695653, array([0.906, 0.948]))
Employee Satisfaction - During Pandemic             : (0.5726036207956922, array([0.427, 0.69 ]))
Employee Satisfaction - Reporting Manager's Support : (0.9267782600741739, array([0.902, 0.947]))
```

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
v1     0
v2     0
v3     0
v4     0
v5     0
v6     0
v7     0
```

```
v42    0
v43    0
v44    0
v45    0
v46    0
dtype: int64
```

In [7]:
```
df["v10"].fillna(0,inplace=True)
df["v19"].fillna("NA",inplace=True)
df["v8"].fillna("NA",inplace=True)
df["v11"].fillna("NA",inplace=True)
df["v22"].fillna("NA",inplace=True)
df["v23"].fillna("NA",inplace=True)
df["v27"].fillna("NA",inplace=True)
df["v28"].fillna("NA",inplace=True)
df["v29"].fillna("NA",inplace=True)
df["v34"].fillna("NA",inplace=True)
df["v35"].fillna("NA",inplace=True)
df["v41"].fillna("NA",inplace=True)
df["v24"].fillna("NA",inplace=True)
```

In [8]:
```
df.isnull().sum()
```

Out[8]:
```
v1     0
v2     0
v3     0
v4     0
v5     0
v6     0
v7     0
v8     0
v9     0
v10    0
v11    0
v12    0
v13    0
v14    0
v15    0
v16    0
v17    0
```

```
v40    0
v41    0
v42    0
v43    0
v44    0
v45    0
v46    0
dtype: int64
```

In [9]:
```
df1 = pd.get_dummies(df, columns = ['v3'])
df1
```

Out[9]:

	v1	v2	v4	v5	v6	v7	v8	v9	v10	v11	...	v39	v40	v41	v42	v43	v44	v45	v46	v3_Female	v3_Male
0	1	Shravya	25	B.Tech	24	Amazon	ROC transportation Specialist	12	500000	5.0	...	1	5	4.0	5	5	5	4	5	1	0
1	2	Venkata Ananda Sai Tunuguntala	23	B.Tech	24	WiSig Networks	Product development Engineer	7	650000	3.0	...	2	5	5.0	5	5	5	5	5	0	1
2	3	REMELLA SATYA RAMANA KUMAR	24	B.Tech	20	Infosys	Senior system Engineer	20	28000	5.0	...	1	5	4.0	5	4	4	3	4	0	1
3	4	R Sai Surya Siva Prasad	23	Masters	12	Accenture	Software Engineer	3	800000	3.0	...	2	3	3.0	2	4	3	2	2	0	1
4	5	AKSHAY CH	25	B.Tech	16	Cigniti tech2logies	Associate engineer	3	400000	4.0	...	1	5	4.0	4	4	4	3	3	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98	99	Leeladhar Reddy Madireddy	25	B.Tech	10	TCS	Asst. System Engineer	10	600000	3.0	...	1	5	4.0	4	3	4	4	3	0	1
99	100	Akhil Maddu	25	B.Tech	43	Accenture	Application Development Senior Analyst	8	800000	3.0	...	2	5	5.0	4	3	4	3	2	0	1
100	101	Mallika	23	B pharm	24	Medico Health Care	Ar Caller	22	250000	3.0	...	1	4	3.0	3	1	1	3	3	1	0
101	102	Sai simha	29	PGDM	4	Tata aig	Associate Management Trainee	4	35000	1.0	...	2	5	5.0	5	5	5	5	5	0	1
102	103	Dheeraj PK	26	Post graduate	60	Mediyea	Executive	48	2800000	2.0	...	1	5	4.0	4	4	4	4	4	0	1

103 rows × 47 columns

103 rows × 47 columns

In [10]:
```
df2= df1.drop(['v2'], axis = 1)
df2
```

Out[10]:

	v1	v4	v5	v6	v7	v8	v9	v10	v11	v12	...	v39	v40	v41	v42	v43	v44	v45	v46	v3_Female	v3_Male
0	1	25	B.Tech	24	Amazon	ROC transportation Specialist	12	500000	5.0	5	...	1	5	4.0	5	5	5	4	5	1	0
1	2	23	B.Tech	24	WiSig Networks	Product development Engineer	7	650000	3.0	5	...	2	5	5.0	5	5	5	5	5	0	1
2	3	24	B.Tech	20	Infosys	Senior system Engineer	20	28000	5.0	5	...	1	5	4.0	5	4	4	3	4	0	1
3	4	23	Masters	12	Accenture	Software Engineer	3	800000	3.0	3	...	2	3	3.0	2	4	3	2	2	0	1
4	5	25	B.Tech	16	Cigniti tech2logies	Associate engineer	3	400000	4.0	5	...	1	5	4.0	4	4	4	3	3	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98	99	25	B.Tech	10	TCS	Asst. System Engineer	10	600000	3.0	3	...	1	5	4.0	4	3	4	4	3	0	1
99	100	25	B.Tech	43	Accenture	Application Development Senior Analyst	8	800000	3.0	3	...	2	5	5.0	4	3	4	3	2	0	1
100	101	23	B pharm	24	Medico Health Care	Ar Caller	22	250000	3.0	2	...	1	4	3.0	3	1	1	3	3	1	0
101	102	29	PGDM	4	Tata aig	Associate Management Trainee	4	35000	1.0	1	...	2	5	5.0	5	5	5	5	5	0	1
102	103	26	Post graduate	60	Mediyea	Executive	48	2800000	2.0	2	...	1	5	4.0	4	4	4	4	4	0	1

103 rows × 46 columns

In [11]:
```
bins = [0,25,50,75,100]

df2['v4'] = pd.cut(df2['v4'], bins)
df2.head()
```

Out[11]:

	v1	v4	v5	v6	v7	v8	v9	v10	v11	v12	...	v39	v40	v41	v42	v43	v44	v45	v46	v3_Female	v3_Male
0	1	(0, 25]	B.Tech	24	Amazon	ROC transportation Specialist	12	500000	5.0	5	...	1	5	4.0	5	5	5	4	5	1	0
1	2	(0, 25]	B.Tech	24	WiSig Networks	Product development Engineer	7	650000	3.0	5	...	2	5	5.0	5	5	5	5	5	0	1

In [11]:
```
bins = [0,25,50,75,100]

df2['v4'] = pd.cut(df2['v4'], bins)
df2.head()
```


Out[11]:

	v1	v4	v5	v6	v7	v8	v9	v10	v11	v12	...	v39	v40	v41	v42	v43	v44	v45	v46	v3_Female	v3_Male
0	1	(0, 25]	B.Tech	24	Amazon	ROC transportation Specialist	12	500000	5.0	5	...	1	5	4.0	5	5	5	4	5	1	0
1	2	(0, 25]	B.Tech	24	WiSig Networks	Product development Engineer	7	650000	3.0	5	...	2	5	5.0	5	5	5	5	5	0	1
2	3	(0, 25]	B.Tech	20	Infosys	Senior system Engineer	20	28000	5.0	5	...	1	5	4.0	5	4	4	3	4	0	1
3	4	(0, 25]	Masters	12	Accenture	Software Engineer	3	800000	3.0	3	...	2	3	3.0	2	4	3	2	2	0	1
4	5	(0, 25]	B.Tech	16	Cigniti tech2logies	Associate engineer	3	400000	4.0	5	...	1	5	4.0	4	4	4	3	3	0	1

5 rows × 46 columns

In [12]:
```
bins = [0,100,200,300,400]

df2['v6'] = pd.cut(df2['v6'], bins)
df2.head()
```

Out[12]:

	v1	v4	v5	v6	v7	v8	v9	v10	v11	v12	...	v39	v40	v41	v42	v43	v44	v45	v46	v3_Female	v3_Male
0	1	(0, 25]	B.Tech	(0, 100]	Amazon	ROC transportation Specialist	12	500000	5.0	5	...	1	5	4.0	5	5	5	4	5	1	0
1	2	(0, 25]	B.Tech	(0, 100]	WiSig Networks	Product development Engineer	7	650000	3.0	5	...	2	5	5.0	5	5	5	5	5	0	1
2	3	(0, 25]	B.Tech	(0, 100]	Infosys	Senior system Engineer	20	28000	5.0	5	...	1	5	4.0	5	4	4	3	4	0	1
3	4	(0, 25]	Masters	(0, 100]	Accenture	Software Engineer	3	800000	3.0	3	...	2	3	3.0	2	4	3	2	2	0	1
4	5	(0, 25]	B.Tech	(0, 100]	Cigniti tech2logies	Associate engineer	3	400000	4.0	5	...	1	5	4.0	4	4	4	3	3	0	1

5 rows × 46 columns

Frequency Distributions - Demographic Data

- Features: v3, v4, v6, v9, v10
- Interpretation of Results

In [13]:
```
df2.v3_Female.value_counts()
df2.v3_Male.value_counts()
```

Out[13]:
```
1    66
0    37
Name: v3_Male, dtype: int64
```

Interpretation:

    From the data there are 37 female and 66 male and the data type is integer

In [14]:
```
df2.v4.value_counts()
```

Out[14]:
```
(0, 25]      77
(25, 50]     25
(50, 75]      1
(75, 100]     0
Name: v4, dtype: int64
```

In [15]:
```
df2.v6.value_counts()
```

Out[15]:
```
(0, 100]      96
(200, 300]     1
(300, 400]     1
(100, 200]     0
Name: v6, dtype: int64
```

In [16]:
```
df2.v9.value_counts()
```

Out[16]:
```
12     13
```

```
310     1
288     1
72      1
27      1
76      1
48      1
Name: v9, dtype: int64
```

In [17]:
```
df2.v10.value_counts()
```

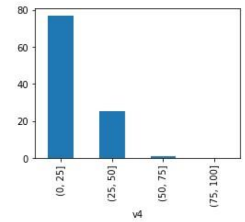
Out[17]:
```
0          8
700000     7
800000     6
600000     6
500000     5
250000     5
15000      4
450000     4
20000      3
1200000    3
650000     3
35000      3
1000000    3
60000      2
400000     2
50000      2
25000      2
40000      2
80000      2
28000      2
45000      2
240000     1
100000     1
70000      1
1900000    1
22345      1
2000000    1
1400000    1
72000      1
200000     1
```

```
Name: v10, dtype: int64
```

In [18]:
```
### Graphical Represention of Numerical Feature
```

In [19]:
```
gd = df2.groupby('v4').size()
gd.plot(kind='bar', figsize=(4,3))
```

Out[19]:
```
<AxesSubplot:xlabel='v4'>
```



In [20]:
```
# Group by Gender (v3) and Overall Satisfaction (v17) of employees with the current organization

ctgos = df2.groupby(['v3_Female', 'v17'])['v3_Female'].count()
gosp  = ctgos.groupby(level=0).apply(lambda x:100 * x / float(x.sum()))
gosp.round()
```

Out[20]:
```
v3_Female  v17
0          1      27.0
           2      15.0
           3      26.0
           4      11.0
           5      21.0
1          1      30.0
           2      16.0
           3      16.0
```

Ln -19:

- The above graphs show the age wise distribution of people
- In the age group of 0 to 25 years there are – 75
- In the age group of 25 to 50 years there are – 20
- In the age group of 50 to 75 years there are – 5
- In the age group of 75 to 100 years there are – 0

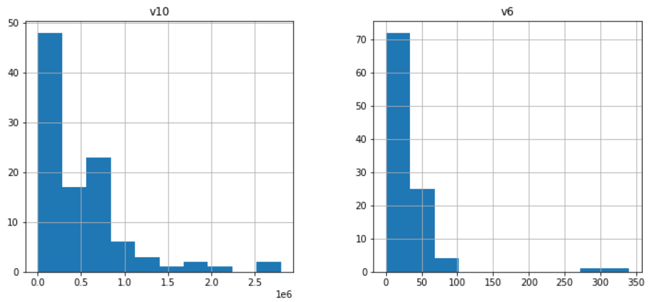
Ln – 21:

- The above histograms represents numerical distribution of total experience and Salary.
- In salary, majority is around 45000 per month and in Experience it is around 70 months

```
Name: v3_Female, dtype: float64
```

In [21]:
```
# Plotting histograms for multiple features together

df.hist(['v10','v6'], figsize=(12,5))
```

Out[21]:
```
array([[<AxesSubplot:title={'center':'v10'}>,
        <AxesSubplot:title={'center':'v6'}>]], dtype=object)
```



In [22]:
```
a = pd.crosstab(index = df2.v3_Female, columns = [df2.v14,df2.v15,df2.v17])
a
```

Out[22]:

v14								1		2	...					4					5
v15	1			2	3		4	5		1	...	2		3	4	5		1	2	3	4
v17	1	2	3	5	5	4	5	5	1	2	...	2	2	4	4	5	1	3	3	3	3
v3_Female																					
0	0	0	0	1	0	1	1	5	2	1	...	2	1	1	0	1	4	1	1	1	1

In [22]:
```
a = pd.crosstab(index = df2.v3_Female, columns = [df2.v14,df2.v15,df2.v17])
a
```

Out[22]:

v14								1		2	...					4					5
v15	1			2	3		4	5		1	...	2		3	4	5		1	2	3	4
v17	1	2	3	5	5	4	5	5	1	2	...	2	2	4	4	5	1	3	3	3	3
v3_Female																					
0	0	0	0	1	0	1	1	5	2	1	...	2	1	1	0	1	4	1	1	1	1
1	1	1	1	0	2	0	0	2	2	0	...	1	1	0	2	0	0	0	1	0	0

2 rows × 45 columns

In [23]:
```
from scipy.stats import chi2_contingency


stat, p, dof, expected = chi2_contingency(a)

alpha = 0.05
print("p value is " + str(p))
if p <= alpha:
    print('Dependent (reject H0)')
else:
    print('Independent (H0 holds true)')
```

```
p value is 0.47487219212325693
Independent (H0 holds true)
```

In [24]:
```
b = pd.crosstab(index = df2.v4, columns = [df2.v14,df2.v15,df2.v17])
b
```

Out[24]:

v14								1		2	...					4					5
v15	1			2	3		4	5		1	...	2		3	4	5		1	2	3	4

```
p value is 0.47487219212325693
Independent (H0 holds true)
```

In [24]:

```python
b = pd.crosstab(index = df2.v4, columns = [df2.v14,df2.v15,df2.v17])
b
```

Out[24]:

v14							1		2	...					4					5
v15	1			2	3		4	5	1	...	2		3	4	5		1	2	3	4
v17	1	2	3	5	5	4	5	5	1	2	...	2	2	4	4	5	1	3	3	3	3
v4																					
(0, 25]	1	1	1	1	1	1	1	5	2	1	...	2	1	0	1	1	3	1	2	1	1
(25, 50]	0	0	0	0	1	0	0	2	2	0	...	1	1	1	1	0	1	0	0	0	0
(50, 75]	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

3 rows × 45 columns

In [25]:

```python
c = pd.crosstab(index = df2.v10, columns = [df2.v14,df2.v15,df2.v17])
c
```

Out[25]:

v14								1		2	...					4					5
v15	1			2	3		4	5		1	...	2		3	4	5		1	2	3	4
v17	1	2	3	5	5	4	5	5	1	2	...	2	2	4	4	5	1	3	3	3	3
v10																					
0	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	1
15000	0	0	1	0	0	0	0	0	0	0	...	0	1	0	0	0	1	0	0	0	0
20000	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	0
22345	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
25000	0	0	0	0	1	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	0

48 rows × 45 columns

In [26]:

```python
d = pd.crosstab(index = df2.v3_Female, columns = [df2.v18,df2.v19,df2.v25,df2.v25])
d
```

Out[26]:

v18						1				2	...	4									5
v19				1	2	3				1	...	5		2		3			4		5
v25	1	2	3	4	3	1	1	2	3	4	...	5	1	3	4	5	2	4	5	4	5
v25	1	2	3	4	3	1	1	2	3	4	...	5	1	3	4	5	2	4	5	4	5
v3_Female																					
0	2	0	2	1	1	1	1	1	1	1	...	2	1	1	3	2	0	2	1	1	6
1	0	1	0	0	0	0	1	1	0	0	...	0	0	0	1	0	1	2	1	1	5

2 rows × 46 columns

In [27]:

```python
e = pd.crosstab(index = df2.v4, columns = [df2.v18,df2.v19,df2.v25,df2.v25])
e
```

Out[27]:

v18						1				2	...	4									5
v19				1	2	3				1	...	5		2		3			4		5
v25	1	2	3	4	3	1	1	2	3	4	...	5	1	3	4	5	2	4	5	4	5
v25	1	2	3	4	3	1	1	2	3	4	...	5	1	3	4	5	2	4	5	4	5
v4																					
(0, 25]	0	1	1	1	1	1	2	1	0	1	...	2	1	0	4	1	1	2	1	2	11
(25, 50]	1	0	1	0	0	0	0	1	1	0	...	0	0	1	0	1	0	2	1	0	0
(50, 75]	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

3 rows × 46 columns

3 rows × 46 columns

In [28]:
```
f = pd.crosstab(index = df2.v10, columns = [df2.v18,df2.v19,df2.v25,df2.v25])
f
```

Out[28]:

v18						1				2	…	4									5
v19				1	2	3				1	…	5		2		3			4		5
v25	1	2	3	4	3	1	1	2	3	4	…	5	1	3	4	5	2	4	5	4	5
v25	1	2	3	4	3	1	1	2	3	4	…	5	1	3	4	5	2	4	5	4	5
v10																					
0	1	0	1	0	0	0	0	1	0	1	…	1	0	0	0	1	1	0	0	0	0
15000	0	0	0	0	0	1	1	0	0	0	…	0	0	0	0	0	0	0	0	0	0
20000	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	1	0	0	0	0	1
22345	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0
25000	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	2
28000	0	0	1	0	0	0	0	0	0	0	…	0	0	0	1	0	0	0	0	0	0
30000	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	1	0
32000	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	1
35000	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	1	0	0	2
40000	0	0	0	0	0	0	0	0	0	0	…	0	0	0	1	0	0	0	0	0	0
45000	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0
50000	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0
60000	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	1	0	0	0
65000	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0
70000	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0

In [29]:
```
g = pd.crosstab(index = df2.v3_Female, columns = [df2.v26,df2.v29])
g
```

Out[29]:

v26					1					2	…				3					4	5
v29	1.0	2.0	3.0	4.0	5.0	1.0	2.0	3.0	4.0	5.0	…	3.0	4.0	5.0	NA	2.0	3.0	4.0	5.0	NA	5.0
v3_Female																					
0	1	2	2	3	0	1	2	2	3	0	…	11	6	4	1	2	7	8	3	1	5
1	1	1	1	1	1	1	3	4	1	1	…	3	1	1	0	1	2	8	0	0	3

2 rows × 22 columns

In [30]:
```
h = pd.crosstab(index = df2.v4, columns = [df2.v26,df2.v29])
h
```

Out[30]:

v26					1					2	…				3					4	5
v29	1.0	2.0	3.0	4.0	5.0	1.0	2.0	3.0	4.0	5.0	…	3.0	4.0	5.0	NA	2.0	3.0	4.0	5.0	NA	5.0
v4																					
(0, 25]	2	3	3	2	0	1	3	6	4	1	…	8	6	4	1	1	6	12	3	0	8
(25, 50]	0	0	0	2	1	1	2	0	0	0	…	6	1	1	0	2	3	4	0	1	0
(50, 75]	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0

3 rows × 22 columns

In [31]:
```
i = pd.crosstab(index = df2.v10, columns = [df2.v26,df2.v29])
i
```

Out[31]:

v26					1					2	…				3					4	5
v29	1.0	2.0	3.0	4.0	5.0	1.0	2.0	3.0	4.0	5.0	…	3.0	4.0	5.0	NA	2.0	3.0	4.0	5.0	NA	5.0

In [32]:

```
#v3 - v30,v33,v34,v35,v36
j = pd.crosstab(index = df2.v3_Female, columns = [df2.v30,df2.v33,df2.v34,df2.v35,df2.v36])
j
```

Out[32]:

v30								1		2	...										5
v33				1	2		3	5		1	...	3			4						5
v34	1.0	2.0	3.0	4.0	2.0	1.0	3.0	5.0	1.0	3.0	...	5.0			5.0	1.0	3.0	4.0			5.0
v35	1.0	2.0	2.0	1.0	3.0	2.0	3.0	5.0	1.0	3.0	...	5.0	2.0	4.0	5.0	5.0	5.0	4.0	4.0		5.0
v36	1	2	4	1	1	2	4	5	1	3	...	5	5	5	5	5	5	4	5	4	5
v3_Female																					
0	3	0	1	0	1	0	0	1	0	0	...	1	1	0	0	1	1	1	1	0	3
1	2	1	0	1	0	1	1	0	1	1	...	0	0	1	2	0	1	0	0	1	1

2 rows × 68 columns

In [33]:

```
#v3 - v30,v33,v34,v35,v36
k = pd.crosstab(index = df2.v4, columns = [df2.v30,df2.v33,df2.v34,df2.v35,df2.v36])
k
```

Out[33]:

v30								1		2	...										5
v33				1	2		3	5		1	...	3			4						5
v34	1.0	2.0	3.0	4.0	2.0	1.0	3.0	5.0	1.0	3.0	...	5.0			5.0	1.0	3.0	4.0			5.0
v35	1.0	2.0	2.0	1.0	3.0	2.0	3.0	5.0	1.0	3.0	...	5.0	2.0	4.0	5.0	5.0	5.0	4.0	4.0		5.0
v36	1	2	4	1	1	2	4	5	1	3	...	5	5	5	5	5	5	4	5	4	5
v4																					
(0, 25]	3	1	1	1	0	1	1	1	0	1	...	0	1	1	2	0	2	1	1	1	4
(25, 50]	1	0	0	0	1	0	0	0	1	0		1	0	0	0	1	0	0	0	0	0

In [34]:

```
#v3 - v30,v33,v34,v35,v36
l = pd.crosstab(index = df2.v10, columns = [df2.v30,df2.v33,df2.v34,df2.v35,df2.v36])
l
```

Out[34]:

v30								1		2	...										5
v33				1	2		3	5		1	...	3			4						5
v34	1.0	2.0	3.0	4.0	2.0	1.0	3.0	5.0	1.0	3.0	...	5.0			5.0	1.0	3.0	4.0			5.0
v35	1.0	2.0	2.0	1.0	3.0	2.0	3.0	5.0	1.0	3.0	...	5.0	2.0	4.0	5.0	5.0	5.0	4.0	4.0		5.0
v36	1	2	4	1	1	2	4	5	1	3	...	5	5	5	5	5	5	4	5	4	5
v10																					
0	2	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
15000	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
20000	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
22345	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
25000	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	2
28000	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
30000	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
32000	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
35000	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	1	0	0	0	1
40000	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	0
45000	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
50000	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
60000	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
65000	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	0

In [35]:
```
#v3 - v37,v38,v39,v40,v41
k = pd.crosstab(index = df2.v3_Female, columns = [df2.v37,df2.v38,df2.v39,df2.v40,df2.v41])
k
```

Out[35]:

v37	1			2						3	...		4								5
v38	4	1	3	4			1	2		3	...		5	1			4				5
v39	1	2	1	2		1	2	2		1	...	1	2	2	1		2		1		2
v40	4	5	5	5	4	5	5	5	3	5	...	5	5	5	5		5		5		5
v41	4.0	2.0	5.0	4.0	3.0	3.0	3.0	3.0	3.0	3.0	...	5.0	5.0	5.0	5.0	4.0	5.0	4.0	5.0	4.0	5.0
v3_Female																					
0	1	0	0	1	0	0	1	1	0	3	...	2	1	1	1	2	0	3	6	1	1
1	0	1	1	0	1	1	0	0	1	1	...	0	0	0	0	0	2	1	1	0	0

2 rows × 49 columns

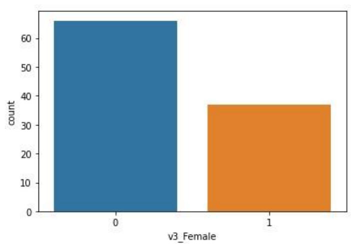
Graphical Represention of Numerical & Categorical Features

1. Plot Categorical Feature - v3 (Gender)
2. Plot Numerical Feature - v10 (Salary)
3. Plot Numerical vs. Numerical Feature - v9 vs. v10 (Experience with current organization and Salary)
4. Plot Numerical vs. Categorical Feature - v3 vs. v10 (Gender vs. Salary)
5. Plot Categorical Feature with multiple categories

In [36]:
```
# countplot() - Plot Categorical feature

sns.countplot(x = "v3_Female", data = df2)
```

Out[36]: <AxesSubplot:xlabel='v3_Female', ylabel='count'>

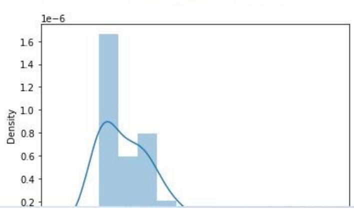
Out[36]: <AxesSubplot:xlabel='v3_Female', ylabel='count'>



In [37]:
```
sns.distplot(df.v10)
```

```
C:\Users\Home\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Pleas
e adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[37]: <AxesSubplot:xlabel='v10', ylabel='Density'>



L 36:

The above graph shows the count of the Gender into male and female categories using bars.

Males and females being 65 and 35 respectively.

L 37:

The above graph depicts the distribution of salary of the employees.

The distribution is represented in terms of density. It ranges from high as 1.6 to low as 0.1 .

L 38:

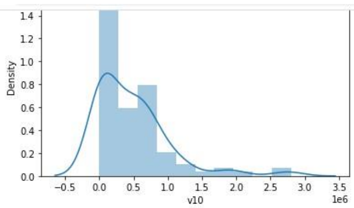
The above scatter plot shows the distribution of salaries with overlaid regression.

Which says that majority of the respondents is fresher.

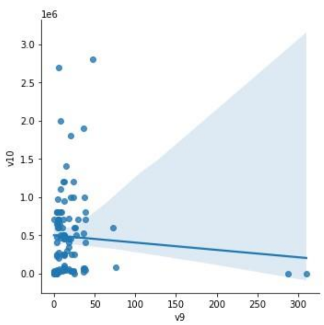
L 39:

The above diagram displays the five number summary of the gender in both male and female. The five figures are

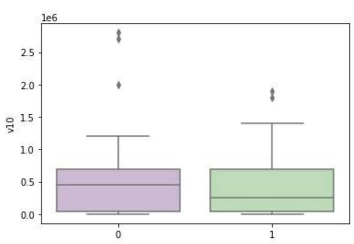
Median, 1st quartile, minimum, maximum, 3rd quartile.



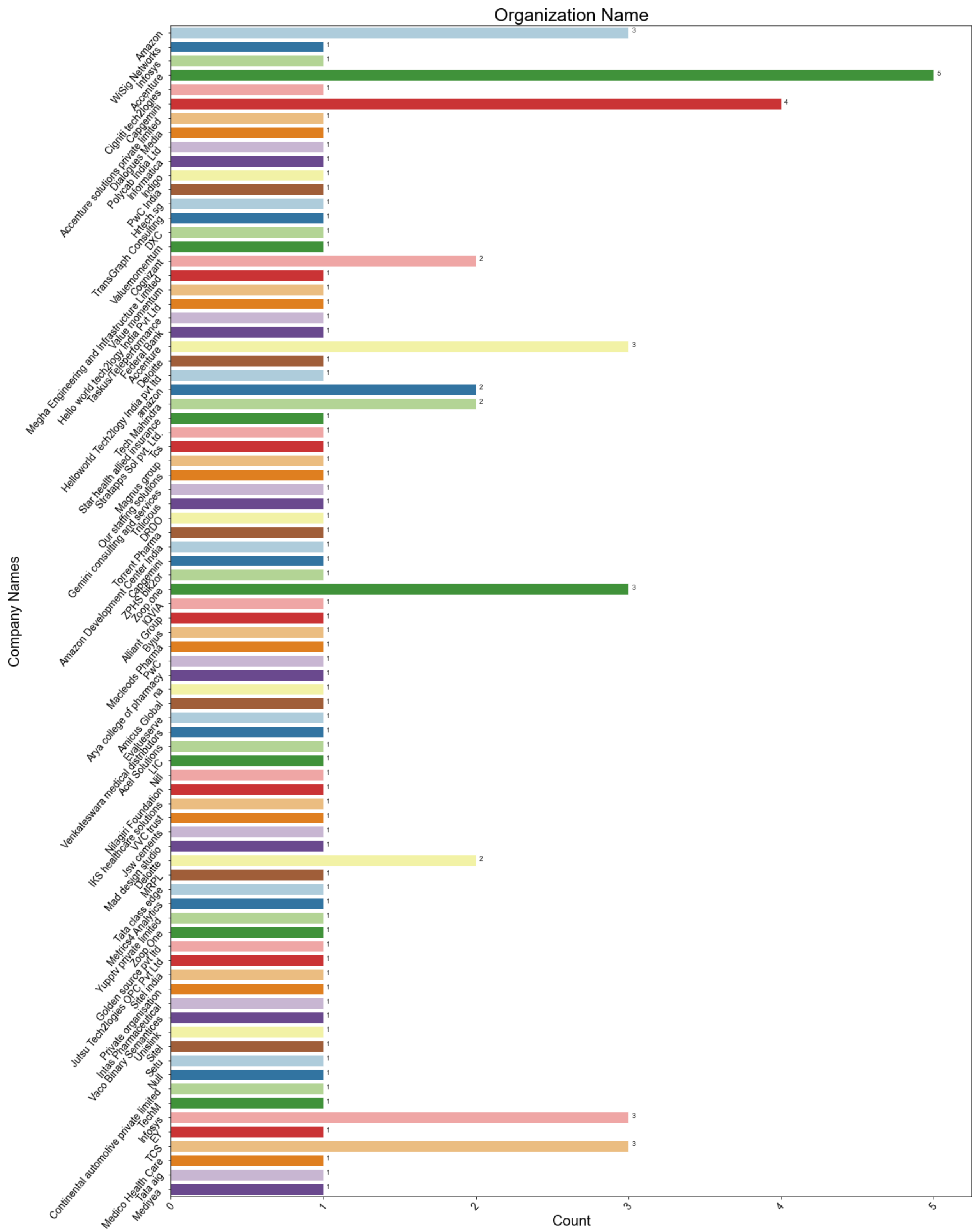
In [38]:
```
sns.lmplot(x = 'v9', y = 'v10', data = df2)
```
Out[38]: <seaborn.axisgrid.FacetGrid at 0x2359243df10>



In [39]:
```
# boxplot() function - Plot Numerical vs. Categorical Features

sns.boxplot(x = 'v3_Female', y = 'v10', data = df2, palette="PRGn")
```
Out[39]: <AxesSubplot:xlabel='v3_Female', ylabel='v10'>



In [40]:
```
# Organization Details (v7) of the Respondents

fig, ax = plt.subplots(figsize =(20,30))
sns.set_style('darkgrid')
plt.fontsize =(35)
ax    = sns.countplot(palette="Paired", ax = ax, y='v7', data = df2)
total = len(df['v7'])
for p in ax.patches:
    percentage ='{:,.0f}'.format(p.get_width())
    width, height =p.get_width(),p.get_height()
    x=p.get_x()+width+0.02
    y=p.get_y()+height/2
    ax.annotate(percentage,(x,y))
ax.grid(False)
plt.xticks(rotation =50,fontsize = 14)
plt.yticks(rotation =50,fontsize = 14)
plt.title('Organization Name', fontsize =25)
plt.xlabel ('Count',fontsize = 20)
plt.ylabel ('Company Names',fontsize = 20)
```
Out[40]: Text(0, 0.5, 'Company Names')



40. A count plot for categorical data which is the “Organization” of the respondents working show the data of 82 companies.

- 5 people are working in Accenture
- 4 people are working in Capgemini
- 3 people a working in Amazon, Zoopone, Infosys and TCS each
- 2 people are working in Tech Mahindra, Deloitte each
- Other organizations have one person working in each

41. A count plot for categorical data which is the “Designation” of the respondents shows the data of 94 different designations.

- Analyst – 3 respondents
- Software engineer, Application development associate, Associate, GET, Associate software engineer, Product manager- 2 respondents each
- Other designation – 1 respondent each

42. This whole graph depicts the “Job satisfaction” of the respondents. Here with the graph, we can interpret that respondents of different organizations’ have different satisfactions level. It can be seen that there is no single satisfaction level (1-5) which has majority respondents. On a whole we can interpret that respondents jobs satisfaction lies between two and four if we wear to calculate the average.

43. It Is seen that majority of both males and females were satisfied regarding the “Overall satisfaction with the organization during the pandemic”

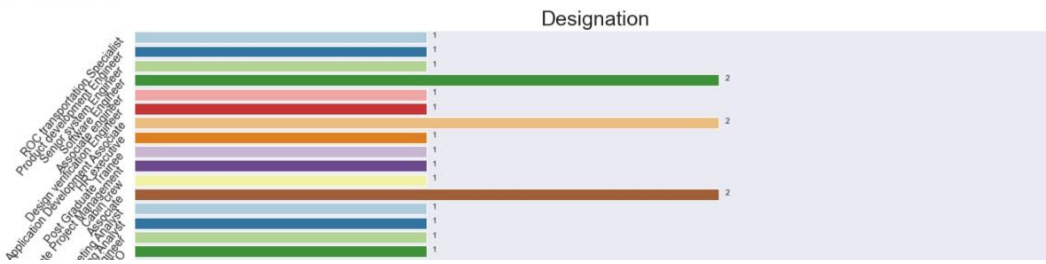
44. Both male and females “Overall satisfaction with organization” ranges from neutral to very satisfied

45. Maximum of the respondents in both males and females opted in the range between 3 to 4 which means neutral to very satisfied for “Overall satisfaction” and “Overall satisfaction during pandemic” combined.

46. KDE plot gives us probability distribution and in both the cases of male and female we can see that the probability is positively skewed.

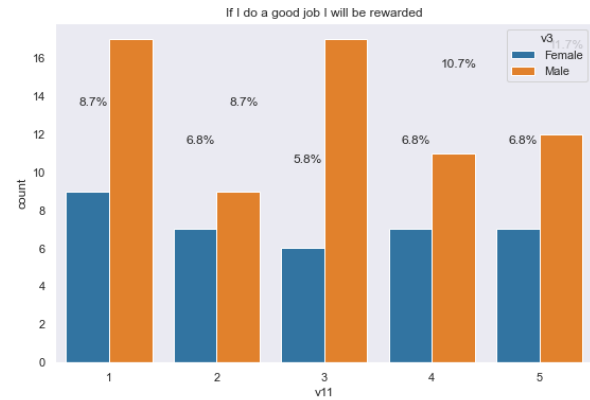
In [41]:
```
# Designation (v8) of the Respondents

fig, ax = plt.subplots(figsize =(20,30))
sns.set_style('darkgrid')
plt.fontsize =(35)
ax    = sns.countplot(palette="Paired", ax = ax, y='v8', data = df2)
total = len(df['v8'])
for p in ax.patches:
    percentage ='{:,.0f}'.format(p.get_width())
    width, height =p.get_width(),p.get_height()
    x=p.get_x()+width+0.02
    y=p.get_y()+height/2
    ax.annotate(percentage,(x,y))
ax.grid(False)
plt.xticks(rotation =50,fontsize = 14)
plt.yticks(rotation =50,fontsize = 14)

plt.title('Designation', fontsize =25)
plt.xlabel ('Count',fontsize = 20)
plt.ylabel ('designation',fontsize = 20)
```

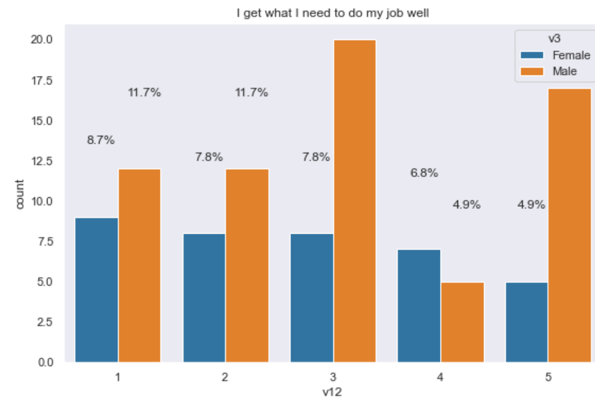
Out[41]:
```
Text(0, 0.5, 'designation')
```

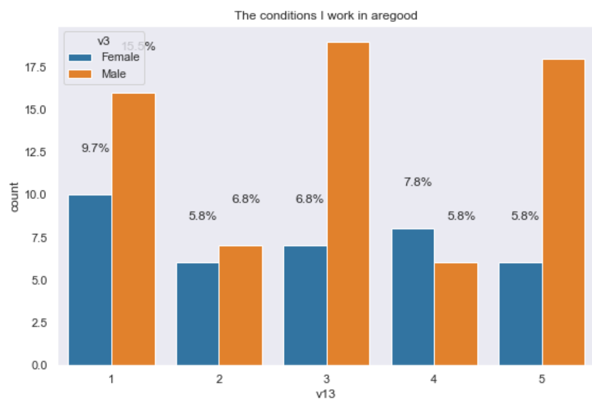


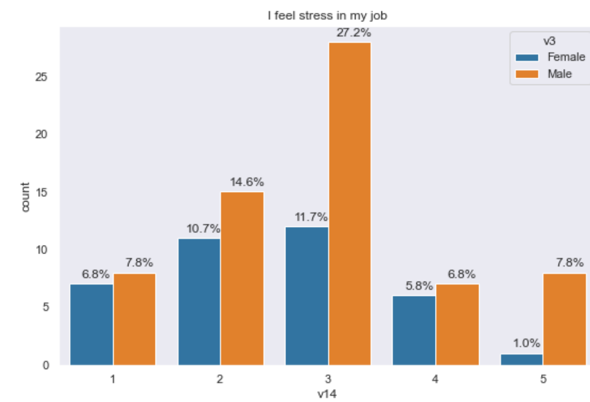
In [42]:
```
# Consolidated graph - General Satisfaction vs. Gender (v3) - v11, v12, v13, v14, v15, v16 vs. v3

sns.set(rc={'figure.figsize':(20,20)})
plt.rcParams['axes.grid'] = False

ax1 = plt.subplot(321)
Rating = [1,2,3,4,5]
hue_order = ["Female","Male"]
ax1 = sns.countplot(x="v11", hue="v3", data=df, order=Rating, hue_order=hue_order, palette="tab10")
total = float(len(df['v3']))
for p in ax1.patches:
    percentage = '{:.1f}%'.format(100 * p.get_height()/total)
    x = p.get_x() + p.get_width() - 0.15
    y = p.get_height() + 4.5
    ax1.annotate(percentage, (x, y),ha='center')

ax2 = plt.subplot(322)
Rating = [1,2,3,4,5]
hue_order = ["Female","Male"]
ax2 = sns.countplot(x="v12", hue="v3", data=df, order=Rating, hue_order=hue_order, palette="tab10")
total = float(len(df['v3']))
for p in ax2.patches:
    percentage = '{:.1f}%'.format(100 * p.get_height()/total)
    x = p.get_x() + p.get_width() - 0.15
    y = p.get_height() + 4.5
    ax2.annotate(percentage, (x, y),ha='center')

ax3 = plt.subplot(323)
Rating = [1,2,3,4,5]
hue_order = ["Female","Male"]
ax3 = sns.countplot(x="v13", hue="v3", data=df, order=Rating, hue_order=hue_order, palette="tab10")
total = float(len(df['v3']))
for p in ax3.patches:
    percentage = '{:.1f}%'.format(100 * p.get_height()/total)
    x = p.get_x() + p.get_width() - 0.15
    y = p.get_height() + 2.5
    ax3.annotate(percentage, (x, y),ha='center')

ax4 = plt.subplot(324)
```

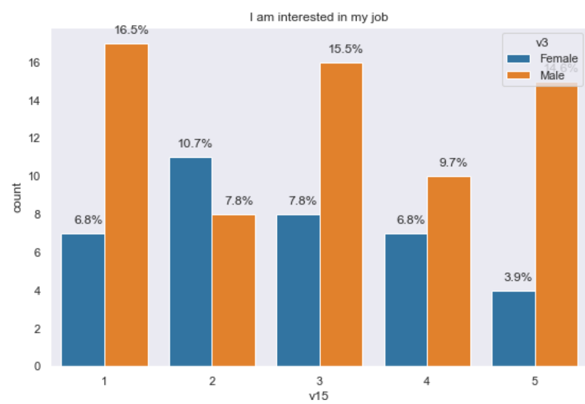
```
ax4 = plt.subplot(324)
Rating = [1,2,3,4,5]
hue_order = ["Female","Male"]
ax4 = sns.countplot(x="v14", hue="v3", data=df, order=Rating, hue_order=hue_order, palette="tab10")
total = float(len(df['v3']))
for p in ax4.patches:
    percentage = '{:.1f}%'.format(100 * p.get_height()/total)
    x = p.get_x() + p.get_width() - 0.15
    y = p.get_height()  + 0.5
    ax4.annotate(percentage, (x, y),ha='center')

ax5 = plt.subplot(325)
Rating = [1,2,3,4,5]
hue_order = ["Female","Male"]
ax5 = sns.countplot(x="v15", hue="v3", data=df, order=Rating, hue_order=hue_order, palette="tab10")
total = float(len(df['v3']))
for p in ax5.patches:
    percentage = '{:.1f}%'.format(100 * p.get_height()/total)
    x = p.get_x() + p.get_width() - 0.15
    y = p.get_height() + 0.5
    ax5.annotate(percentage, (x, y),ha='center')

ax6 = plt.subplot(326)
Rating = [1,2,3,4,5]
hue_order = ["Female","Male"]
ax6 = sns.countplot(x="v16", hue="v3", data=df, order=Rating, hue_order=hue_order, palette="tab10")
total = float(len(df['v3']))
for p in ax6.patches:
    percentage = '{:.1f}%'.format(100 * p.get_height()/total)
    x = p.get_x() + p.get_width() - 0.15
    y = p.get_height() + 0.5
    ax6.annotate(percentage, (x, y),ha='center')

ax1.title.set_text('If I do a good job I will be rewarded')
ax2.title.set_text('I get what I need to do my job well')
ax3.title.set_text('The conditions I work in aregood')
ax4.title.set_text('I feel stress in my job')
ax5.title.set_text('I am interested in my job')
ax6.title.set_text('Management makes wise decisions')
```

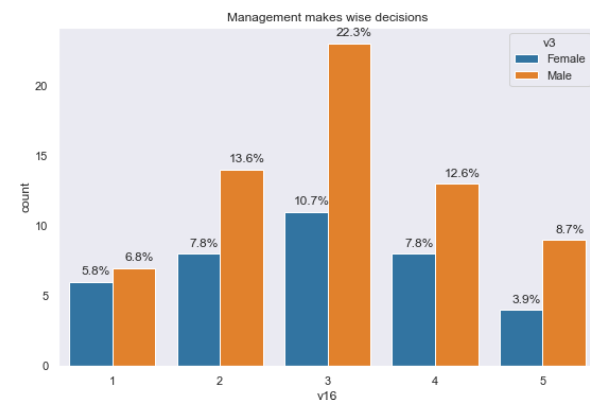

If I do a good job I will be rewarded
count
v11
v3
Female
Male


I get what I need to do my job well
count
v12
v3
Female
Male


The conditions I work in aregood
count
v13
v3
Female
Male


I feel stress in my job
count
v14
v3
Female
Male

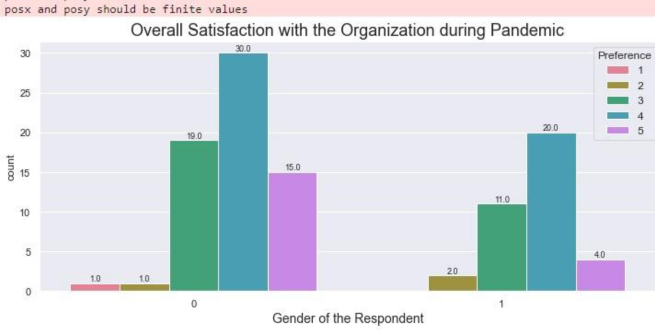

I am interested in my job
count
v15
v3
Female
Male


Management makes wise decisions
count
v16
v3
Female
Male

```
In [43]: # Overall satisfaction of the employee during pandemic against gender of the employee

plt.figure(figsize=(12,5))
sns.set(style="darkgrid")
ax = sns.countplot(x="v3_Female", hue='v37', data=df2, palette="husl")
plt.title("Overall Satisfaction with the Organization during Pandemic",fontsize=18)
plt.xlabel('Gender of the Respondent',fontsize=14)
for rect in ax.patches:
    ax.text (rect.get_x() + rect.get_width()  / 2,rect.get_height()+ 0.25,rect.get_height(),horizontalalignment='center', fontsize = 9)
ax.legend(fontsize = 12, bbox_to_anchor= (1.01, 1), title="Preference", title_fontsize = 12)
```

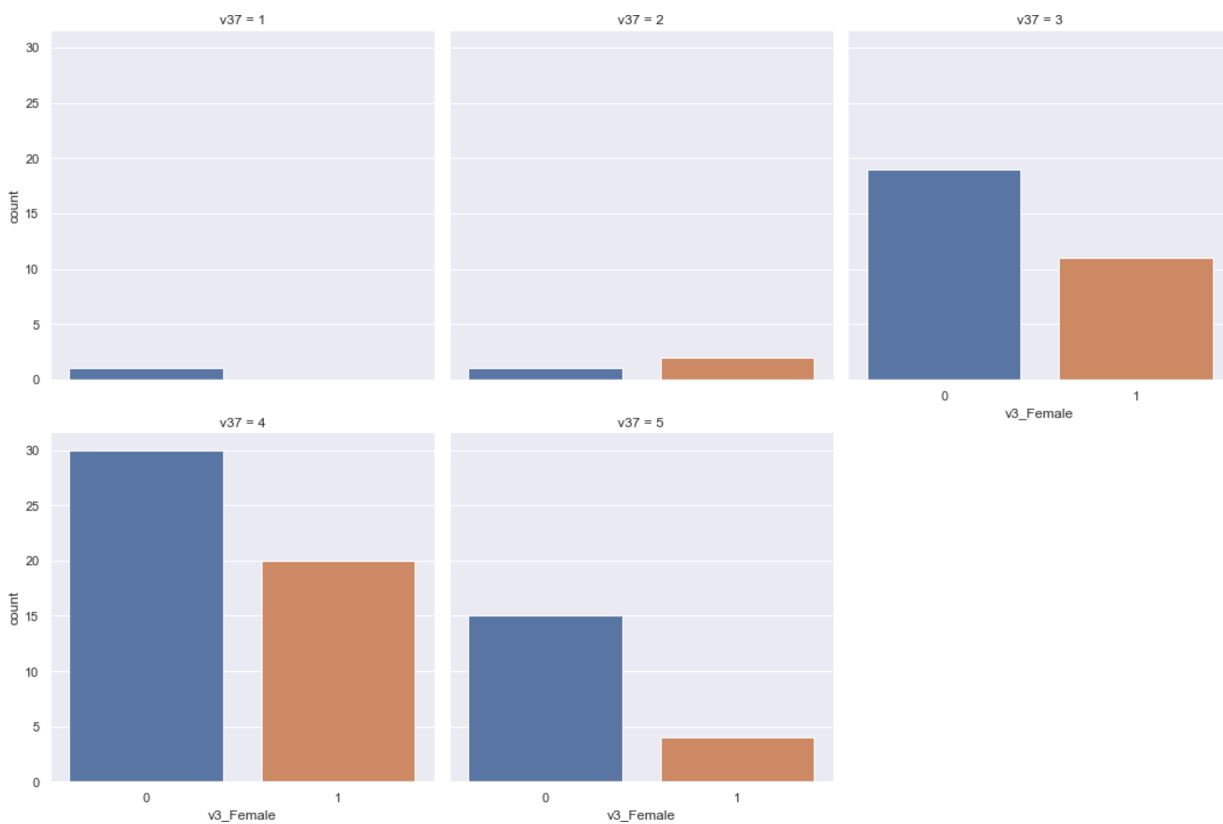
```
Out[43]: <matplotlib.legend.Legend at 0x23593f99820>
```

```
posx and posy should be finite values
posx and posy should be finite values
```




```
In [44]: # Overall Satisfaction vs. Gender

sns.factorplot("v3_Female", col = "v37", col_wrap = 3, data = df2,kind = "count")
plt.show()
```

```
C:\Users\Home\anaconda3\lib\site-packages\seaborn\categorical.py:3717: UserWarning: The `factorplot` function has been renamed to `catplot`. The original name will be rem
oved in a future release. Please update your code. Note that the default `kind` in `factorplot` (`'point'`) has changed `'strip'` in `catplot`.
  warnings.warn(msg)
C:\Users\Home\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid pos
itional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

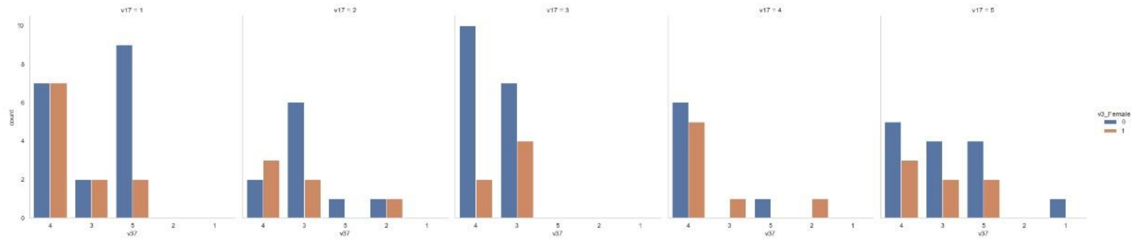


In [45]:

```
# Gender-wise, Overall Satisfaction and Satisfaction during Pandemic

sns.set(style="whitegrid")
plt.rcParams['axes.grid'] = False
sns.catplot(x="v37", hue="v3_Female", col="v17",data=df2, order = df['v37'].value_counts().index,kind="count",height=6, aspect=0.9)
```

Out[45]:

```
<seaborn.axisgrid.FacetGrid at 0x235944c66d0>
```

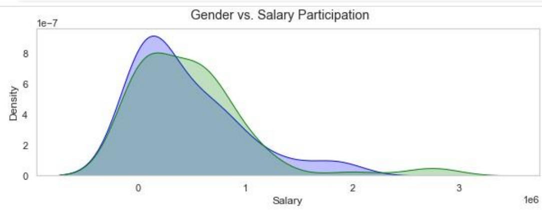


In [46]:

```
# Gender (0-Male, 1-Female) vs. salary

fig, ax = plt.subplots(figsize=(10,3))

sns.kdeplot(df2[df2["v3_Female"]==1]["v10"], shade=True, color="blue", ax=ax)
sns.kdeplot(df2[df2["v3_Female"]==0]["v10"], shade=True, color="green", ax=ax)

ax.set_xlabel("Salary")
ax.set_ylabel("Density")
fig.suptitle("Gender vs. Salary Participation")
```

Out[46]:

```
Text(0.5, 0.98, 'Gender vs. Salary Participation')
```



In [47]:

```
df2.describe(include='all')
```

Out[47]:

	v1	v4	v5	v6	v7	v8	v9	v10	v11	v12	...	v39	v40	v41	v42	v43	v44	v45	v46	v3_
count	103.000000	103	103	98	103	103	103.000000	1.030000e+02	103.0	103.000000	...	103.000000	103.000000	103.0	103.000000	103.000000	103.000000	103.000000	103.000000	103
unique	NaN	3	31	3	82	94	NaN	NaN	6.0	NaN	...	NaN	NaN	5.0	NaN	NaN	NaN	NaN	NaN	
top	NaN	(0, 25]	B.Tech	(0, 100]	Accenture	Analyst	NaN	NaN	1.0	NaN	...	NaN	NaN	4.0	NaN	NaN	NaN	NaN	NaN	
freq	NaN	77	31	96	5	3	NaN	NaN	26.0	NaN	...	NaN	NaN	54.0	NaN	NaN	NaN	NaN	NaN	
mean	52.000000	NaN	NaN	NaN	NaN	NaN	20.932039	4.782742e+05	NaN	2.941748	...	1.339806	4.815534	NaN	3.679612	3.669903	3.660194	3.708738	3.621359	0
std	29.877528	NaN	NaN	NaN	NaN	NaN	41.785930	5.427169e+05	NaN	1.413002	...	0.475959	0.537744	NaN	1.095567	1.069986	1.107491	1.063024	1.103617	0
min	1.000000	NaN	NaN	NaN	NaN	NaN	0.000000	0.000000e+00	NaN	1.000000	...	1.000000	3.000000	NaN	1.000000	1.000000	1.000000	1.000000	1.000000	0
25%	26.500000	NaN	NaN	NaN	NaN	NaN	6.000000	4.000000e+04	NaN	2.000000	...	1.000000	5.000000	NaN	3.000000	3.000000	3.000000	3.000000	3.000000	0
50%	52.000000	NaN	NaN	NaN	NaN	NaN	12.000000	4.000000e+05	NaN	3.000000	...	1.000000	5.000000	NaN	4.000000	4.000000	4.000000	4.000000	4.000000	0
75%	77.500000	NaN	NaN	NaN	NaN	NaN	24.000000	7.000000e+05	NaN	4.000000	...	2.000000	5.000000	NaN	5.000000	4.000000	4.500000	5.000000	4.000000	1
max	103.000000	NaN	NaN	NaN	NaN	NaN	310.000000	2.800000e+06	NaN	5.000000	...	2.000000	5.000000	NaN	5.000000	5.000000	5.000000	5.000000	5.000000	1

In [48]:
```
df2.describe(include=['object'])
```

Out[48]:

	v5	v7	v8	v11	v22	v23	v24	v27	v28	v29	v34	v35	v41
count	103	103	103	103.0	103.0	103.0	103.0	103.0	103.0	103.0	103.0	103.0	103.0
unique	31	82	94	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	5.0
top	B.Tech	Accenture	Analyst	1.0	3.0	5.0	3.0	4.0	3.0	3.0	3.0	3.0	4.0
freq	31	5	3	26.0	28.0	34.0	36.0	29.0	36.0	32.0	32.0	31.0	54.0

In [49]:
```
df.describe(include=['int64'])
```

Out[49]:

	v1	v4	v6	v9	v10	v12	v13	v14	v15	v16	...	v36	v37	v38	v39	v40	v4
count	103.000000	103.00000	103.000000	103.000000	1.030000e+02	103.000000	103.000000	103.000000	103.000000	103.000000	...	103.000000	103.000000	103.000000	103.000000	103.000000	103.00000
mean	52.000000	25.23301	28.844660	20.932039	4.782742e+05	2.941748	2.970874	2.757282	2.883495	2.990291	...	3.398058	3.805825	3.757282	1.339806	4.815534	3.67961
std	29.877528	4.23386	44.597933	41.785930	5.427169e+05	1.413002	1.491521	1.124295	1.423204	1.200450	...	1.231449	0.805109	0.974841	0.475959	0.537744	1.09556
min	1.000000	21.00000	0.000000	0.000000	0.000000e+00	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	3.000000	1.00000
25%	26.500000	23.00000	7.500000	6.000000	4.000000e+04	2.000000	1.500000	2.000000	2.000000	2.000000	...	3.000000	3.000000	3.000000	1.000000	5.000000	3.00000
50%	52.000000	24.00000	23.000000	12.000000	4.000000e+05	3.000000	3.000000	3.000000	3.000000	3.000000	...	4.000000	4.000000	4.000000	1.000000	5.000000	4.00000
75%	77.500000	25.50000	36.000000	24.000000	7.000000e+05	4.000000	4.000000	3.000000	4.000000	4.000000	...	4.000000	4.000000	4.000000	2.000000	5.000000	5.00000
max	103.000000	52.00000	340.000000	310.000000	2.800000e+06	5.000000	5.000000	5.000000	5.000000	5.000000	...	5.000000	5.000000	5.000000	2.000000	5.000000	5.00000

8 rows × 31 columns

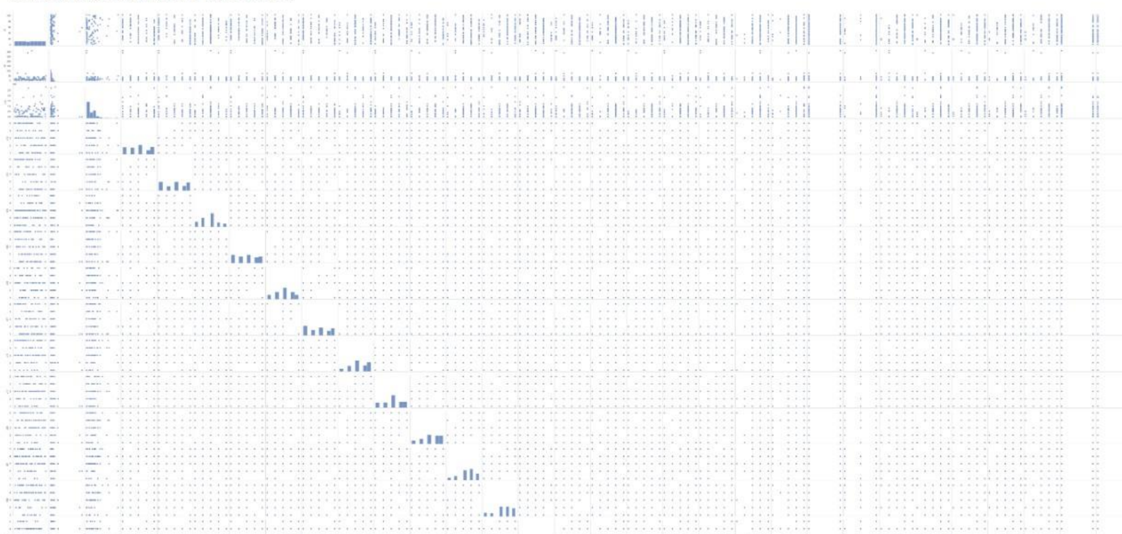
In [50]:
```
df2.corr()
```

Out[50]:

	v1	v9	v10	v12	v13	v14	v15	v16	v17	v18	...	v38	v39	v40	v42	v43	v44	v45	
v1	1.000000	0.012313	0.186910	-0.267061	-0.232982	0.165777	-0.207967	-0.118358	-0.192966	0.034817	...	-0.075736	0.051707	0.025629	0.061101	0.032507	0.045628	0.000309	-0.00:

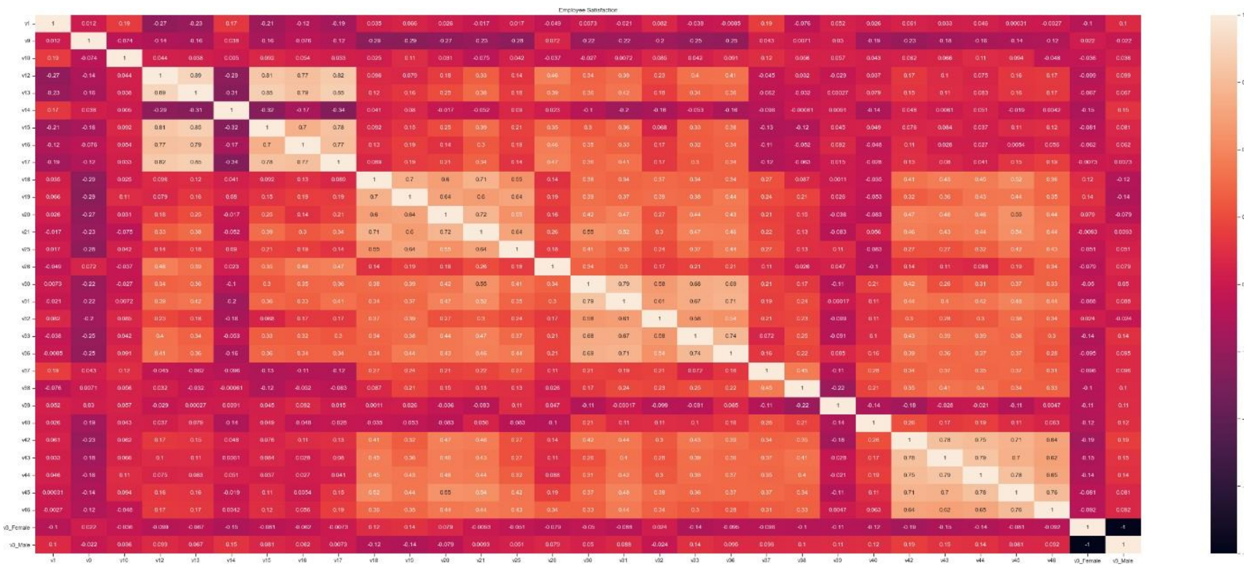
In [51]:
```
sns.pairplot(df2)
```

Out[51]:
```
<seaborn.axisgrid.PairGrid at 0x235947cc1c0>
```



In [52]:
```
# heatmap() - Measures association between muliptle features

plt.figure(figsize=(50,20))
sns.set_style('ticks')

sns.heatmap(df2.corr(), annot=True)

plt.title('Employee Satisfaction')
plt.show()
```

Employee Satisfaction




```
In [53]: df2.corr().style.background_gradient(cmap='coolwarm')
```

	v1	v9	v10	v12	v13	v14	v15	v16	v17
v1	1.000000	0.012313	0.186910	-0.267061	-0.232982	0.165777	-0.207967	-0.118358	-0.192966
v9	0.012313	1.000000	-0.073977	-0.137055	-0.162055	0.037626	-0.164330	-0.076432	-0.123063
v10	0.186910	-0.073977	1.000000	0.044136	0.037815	0.005039	0.092087	0.053843	0.033425
v12	-0.267061	-0.137055	0.044136	1.000000	0.887695	-0.286696	0.805871	0.768377	0.823262
v13	-0.232982	-0.162055	0.037815	0.887695	1.000000	-0.314118	0.848195	0.788317	0.848346
v14	0.165777	0.037626	0.005039	-0.286696	-0.314118	1.000000	-0.318072	-0.168835	-0.335446
v15	-0.207967	-0.164330	0.092087	0.805871	0.848195	-0.318072	1.000000	0.699412	0.784802
v16	-0.118358	-0.076432	0.053843	0.768377	0.788317	-0.168835	0.699412	1.000000	0.765716
v17	-0.192966	-0.123063	0.033425	0.823262	0.848346	-0.335446	0.784802	0.765716	1.000000
v18	0.034817	-0.285029	0.025218	0.095532	0.115516	0.041094	0.091511	0.131910	0.089352
v19	0.066401	-0.291058	0.111598	0.078712	0.156321	0.080326	0.150871	0.193350	0.187499
v20	0.025961	-0.271699	0.030506	0.180832	0.252995	-0.017309	0.245755	0.136766	0.210181
v21	-0.017212	-0.229910	-0.075004	0.325610	0.381220	-0.052106	0.387591	0.303060	0.339308
v25	0.017065	-0.280065	0.041831	0.142232	0.180442	0.090157	0.205264	0.192347	0.141298
v26	-0.049270	0.072357	-0.037476	0.461742	0.390861	0.022615	0.351589	0.456561	0.468860
v30	0.007332	-0.222574	-0.027049	0.341398	0.357543	-0.100540	0.295945	0.353092	0.361075
v31	-0.020747	-0.216777	0.007239	0.385844	0.418639	-0.202173	0.360050	0.331543	0.405878
v32	0.082491	-0.196217	0.085305	0.226747	0.182516	-0.160176	0.067908	0.168912	0.166963
v33	-0.037635	-0.246451	0.042130	0.404311	0.338411	-0.053405	0.326455	0.319227	0.295448
v36	-0.008527	-0.252868	0.091187	0.407858	0.364000	-0.156129	0.356762	0.340868	0.337199
v37	0.190742	0.043316	0.121511	-0.044512	-0.061906	-0.095901	-0.131166	-0.113551	-0.118357

	v18	v19	v20	v21	v25	v26	v30	v31	v32	v33
v1	0.034817	0.066401	0.025961	-0.017212	0.017065	-0.049270	0.007332	-0.020747	0.082491	-0.037635
v9	-0.285029	-0.291058	-0.271699	-0.229910	-0.280065	0.072357	-0.222574	-0.216777	-0.196217	-0.246451
v10	0.025218	0.111598	0.030506	-0.075004	0.041831	-0.037476	-0.027049	0.007239	0.085305	0.042130
v12	0.095532	0.078712	0.180832	0.325610	0.142232	0.461742	0.341398	0.385844	0.226747	0.404311
v13	0.115516	0.156321	0.252995	0.381220	0.180442	0.390861	0.357543	0.418639	0.182516	0.338411
v14	0.041094	0.080326	-0.017309	-0.052106	0.090157	0.022615	-0.100540	-0.202173	-0.160176	-0.053405
v15	0.091511	0.150871	0.245755	0.387591	0.205264	0.351589	0.295945	0.360050	0.067908	0.326455
v16	0.131910	0.193350	0.136766	0.303060	0.192347	0.456561	0.353092	0.331543	0.168912	0.319227
v17	0.089352	0.187499	0.210181	0.339308	0.141298	0.468860	0.361075	0.405878	0.166963	0.295448
v18	1.000000	0.698130	0.604912	0.710504	0.547123	0.138302	0.379357	0.344257	0.374273	0.344822
v19	0.698130	1.000000	0.637029	0.599776	0.635791	0.193594	0.391614	0.371969	0.388257	0.381185
v20	0.604912	0.637029	1.000000	0.722293	0.546613	0.164153	0.416499	0.471978	0.267951	0.436315
v21	0.710504	0.599776	0.722293	1.000000	0.637085	0.255993	0.552091	0.522350	0.303397	0.469271
v25	0.547123	0.635791	0.546613	0.637085	1.000000	0.179766	0.409467	0.347456	0.241425	0.372244
v26	0.138302	0.193594	0.164153	0.255993	0.179766	1.000000	0.340181	0.299381	0.172130	0.213658
v30	0.379357	0.391614	0.416499	0.552091	0.409467	0.340181	1.000000	0.791200	0.579483	0.679186
v31	0.344257	0.371969	0.471978	0.522350	0.347456	0.299381	0.791200	1.000000	0.608420	0.674450
v32	0.374273	0.388257	0.267951	0.303397	0.241425	0.172130	0.579483	0.608420	1.000000	0.578791
v33	0.344822	0.381185	0.436315	0.469271	0.372244	0.213658	0.679186	0.674450	0.578791	1.000000
v36	0.338397	0.440139	0.433269	0.464163	0.440068	0.211333	0.692117	0.706756	0.539784	0.736326
v37	0.269609	0.243145	0.205775	0.222837	0.270244	0.114559	0.210159	0.190303	0.206695	0.071786

L 51:

The above graph shows the multiple pair wise bi-variate distribution of the dataset.

L 52:

The above graph visually depicts the employee satisfaction level. Where the red signifies the highest level of satisfaction.

L 53:

The above color map shows the distribution of the all the variables taken in the study.

simple linear regression

In [54]:
```
df_data = pd.read_excel(r"C:\Users\Home\Downloads\Python for Datascience and ML\Final\Team3_Dataset.xlsx")
df_data
```

Out[54]:

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	...	v37	v38	v39	v40	v41	v42	v43	v44	v45	v46
0	1	Shravya	Female	25	B.Tech	24	Amazon	ROC transportation Specialist	12	500000	...	4	5	1	5	4.0	5	5	5	4	5
1	2	Venkata Ananda Sai Tunuguntala	Male	23	B.Tech	24	WiSig Networks	Product development Engineer	7	650000	...	4	5	2	5	5.0	5	5	5	5	5
2	3	REMELLA SATYA RAMANA KUMAR	Male	24	B.Tech	20	Infosys	Senior system Engineer	20	28000	...	4	4	1	5	4.0	5	4	4	3	4
3	4	R Sai Surya Siva Prasad	Male	23	Masters	12	Accenture	Software Engineer	3	800000	...	3	3	2	3	3.0	2	4	3	2	2
4	5	AKSHAY CH	Male	25	B.Tech	16	Cigniti tech2logies	Associate engineer	3	400000	...	4	4	1	5	4.0	4	4	4	3	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98	99	Leeladhar Reddy Madireddy	Male	25	B.Tech	10	TCS	Asst. System Engineer	10	600000	...	4	3	1	5	4.0	4	3	4	4	3
99	100	Akhil Maddu	Male	25	B.Tech	43	Accenture	Application Development Senior Analyst	8	800000	...	4	3	2	5	5.0	4	3	4	3	2
100	101	Mallika	Female	23	B pharm	24	Medico Health Care	Ar Caller	22	250000	...	3	1	1	4	3.0	3	1	1	3	3
101	102	Sai simha	Male	29	PGDM	4	Tata aig	Associate Management Trainee	4	35000	...	3	5	2	5	5.0	5	5	5	5	5
102	103	Dheeraj PK	Male	26	Post graduate	60	Mediyea	Executive	48	2800000	...	5	5	1	5	4.0	4	4	4	4	4

103 rows × 46 columns

In [55]:
```
import statsmodels.api as sm

X = sm.add_constant(df_data['v4'])
```

In [55]:
```
import statsmodels.api as sm

X = sm.add_constant(df_data['v4'])
Y = df_data['v10']

slm = sm.OLS(Y,X).fit()
print(slm.params)
slm.summary2()
```

```
const    384708.944808
v4         3708.050667
dtype: float64
```

Out[55]:

Model:	OLS	Adj. R-squared:	-0.009
Dependent Variable:	v10	AIC:	3015.3049
Date:	2022-12-09 14:53	BIC:	3020.5743
No. Observations:	103	Log-Likelihood:	-1505.7
Df Model:	1	F-statistic:	0.08459
Df Residuals:	101	Prob (F-statistic):	0.772
R-squared:	0.001	Scale:	2.9721e+11

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	384708.9448	326163.0235	1.1795	0.2410	-262310.7570	1031728.6466
v4	3708.0507	12749.5364	0.2908	0.7718	-21583.6000	28999.7014

Omnibus:	54.717	Durbin-Watson:	1.828
Prob(Omnibus):	0.000	Jarque-Bera (JB):	167.742
Skew:	1.935	Prob(JB):	0.000
Kurtosis:	7.910	Condition No.:	156

```
In [56]: # Multiple Linear Regression - Salary (v10) vs. Age (v4), Total Experience (v6), Experience with Current Organization (v9)

from sklearn.model_selection import train_test_split
X = df_data[['v4','v6','v9']]        # Input/ independent features
Y = df_data['v10']              # Dependent feature

from sklearn.model_selection import train_test_split
trainX, testX, trainY, testY = train_test_split(X , Y, train_size = 0.8, random_state = 100)  # Split the dataset into training (80) and validation (20) sub-sets
print("Input attributes X - Train dataset   :", trainX.shape)
print("Input attributes X - Test dataset    :", testX.shape)
print("Output attribute y - Train dataset   :", trainY.shape)
print("Output attribute y - Test dataset    :", testY.shape)

from sklearn.linear_model import LinearRegression
mlr = LinearRegression().fit(trainX,trainY)
print("Intercept is                         :",mlr.intercept_)     # intercept_ gives Y-intercept value
print("Coefficients of the input features   :",mlr.coef_)

predict_test = mlr.predict(testX)                                # Predict for test dataset
df = pd.DataFrame({'Actual': testY, 'Predicted': predict_test})  # Display Actual against Predicted values
df.head()

from sklearn.metrics import r2_score, mean_squared_error, max_error, mean_absolute_error
print("Accuracy of the model - R square              = ",mlr.score(testX,testY))                # Model Diagnostics
print('Maximum error between original & predicted data = ', max_error(testY, predict_test))
print('Mean Absolute Error (MAE)                     = ', mean_absolute_error(testY, predict_test))
print('Mean squared error (MSE)                      = ', mean_squared_error(testY, predict_test))
print('Root mean squared error (RMSE)                = ', np.sqrt(mean_squared_error(testY, predict_test)))

Input attributes X - Train dataset   : (82, 3)
Input attributes X - Test dataset    : (21, 3)
Output attribute y - Train dataset   : (82,)
Output attribute y - Test dataset    : (21,)
Intercept is                         : -280161.44704078685
Coefficients of the input features   : [ 29998.34409707  11222.4365659  -13866.66417943]
Accuracy of the model - R square              =  -0.5330961574092934
Maximum error between original & predicted data =  879067.1849870207
Mean Absolute Error (MAE)                     =  387062.71922612964
Mean squared error (MSE)                      =  205151210157.59677
Root mean squared error (RMSE)                =  452936.2098106054
```

- We are doing a simple linear regression of **Age** and **Salary**
- **R-Squared:** Lies between 0 & 1. Higher R-Squared indicates better fit. Ideal value of R2 should be 1. In our Regression analysis R-Square is 0.001 which says it is not a good fit
- **Adjusted R^2:** Should be a bit less than the R-Squared value (1). Ours is - 0.009 which is less than R-Squared so it ideal
- **Omnibus (Checks for Normality assumption):** Test of skewness & kurtosis of the residual. We're looking at the distribution of the residual – Null Hypothesis - Errors are normally distributed. Value close to zero would indicate normal distribution. Since ours is 54.717 it is not normally distributed.
- **Prob (Omnibus)**: Performs a statistical test indicating the probability that the residuals are normally distributed. If Prob (Omnibus) is close to 1 then the residuals are normally distributed, satisfying OLS assumption. The value of our Prob (Omnibus) is 0 which says residuals re not normally distributed and it does not satisfy OLS assumption.
- **Skew (Checks for Normality assumption):** Prefer something close to 0, indicating residual distribution is normal - This value also drives the Omnibus. Skew is 1.935 indicating residual distribution is not normal.
- **Kurtosis – Measure of "tailedness":** Higher peaks lead to greater Kurtosis. Greater Kurtosis means tighter clustering of residuals around zero, implying a better model with few outliers. Kurtosis is 7.910
- **Cond. No.** - If condition number is greater than thirty (30), then the regression may have multicollinearity. Condition no. is 156 which means regression has multicollinearity.

Questionnaire

A Study to Measure the Satisfaction Levels of Employees with Changing Workplace

Name of the Respondent	
Gender	
Age	
Qualification	
Total Experience (in months)	
Organization Name	
Designation	
Experience with Current Organization (in months)	
Salary	

1. Please express your level of agreement/ disagreement for the following statements:
   (1 – Strongly Agree, 2 – Agree, 3 – Neutral, 4 – Disagree; 5 – Strongly Disagree)

	1	2	3	4	5
If I do a good job I will be rewarded					
I get what I need to do my job well					
The conditions I work in are good					
I feel stress in my job					
I am interested in my job					
Management makes wise decisions					
I am proud to work for this company					

2. Please express your level of satisfaction/ dissatisfaction with the employee benefits & compensation during pandemic:
   (1 – Poor, 2 – Fair, 3 – Good, 4 – Very Good; 5 – Excellent)

	1	2	3	4	5
Health Care					
Vacation Time					
Training Availability					
Performance Reviews					
Day Care					
Parking					
Retirement Plan					
Salary/Wages					

3. How often do you undergo the following review processes:

   (1 – Weekly, 2 – Monthly, 3 – Quarterly, 4 – Annually; 5 – Never)

	1	2	3	4	5
Personal Performance					
Departmental Performance					
Company Performance					
Career Goals					

4. Please express your level of agreement or disagreement for the following statements:

(1 – Strongly Disagree, 2 – Disagree, 3 – Neutral, 4 – Agree; 5 – Strongly Agree)

	1	2	3	4	5
Company tries to create an exciting work environment					
I receive enough opportunity to interact with other employees on a formal level					
I receive enough opportunity to interact with other employees on an informal level					
I feel a sense of completion with my job					
I have enough freedom in my position to take independent action when needed					
I have enough freedom in my position to do what is right for the customer					
I have a clear path for career advancement					

5. Overall how satisfied are you with your position at this company during pandemic?
   - □ Extremely Dissatisfied
   - □ Dissatisfied
   - □ Neutral
   - □ Satisfied
   - □ Extremely Satisfied

6. How flexible is the company with respect to your family responsibilities?
   - □ Extremely Inflexible
   - □ Inflexible
   - □ Neutral
   - □ Flexible
   - □ Extremely Flexible

7. Do you take part in your company's flexi-time program?
   - □ Yes
   - □ No

8. Have you ever observed or experienced any of the following forms of discrimination or harassment at this company?
   - □ Racial discrimination
   - □ Sexual orientation discrimination
   - □ Gender discrimination
   - □ Age discrimination
   - □ None of the above

9. My job requirements are clear.
   - □ Extremely Dissatisfied
   - □ Dissatisfied
   - □ Neutral
   - □ Satisfied
   - □ Extremely Satisfied

10. How would you rate your reporting manager in each of the following areas during work from home?

(1 – Poor, 2 – Fair, 3 – Good, 4 – Very Good; 5 – Excellent)

	1	2	3	4	5
Communication					
Planning and organizing					
Directing and coordinating					
Job/Technical knowledge					
Interpersonal relationship					

Conclusion

The analysis of workplace conditions during the pandemic is analyzed and following are the inferences:

- Gender and Qualifications doesn’t impact on designation as during the chi2_contingency test the p value was more than 0.05. So, they are independent variables. The gender and qualification of a person doesn’t impact on the designation of a person because a person’s skill or capability can’t be measured from his/her gender or the level of their qualification.

- Experience in an organization can impact on designation. It can indicate how much experience you had gained and how much skills you must lead a team. Experience put impact on designation as more experience you get a better designation in a company. The work you have done in the past and the projects you have undertaken are particular to you alone. A relevant experience makes you stand out from the rest of the rat pack even if you choose to go for a formal education later.

- Rewards for good work put impact on interest in job as the more rewards a person gets for his/her work the more interest he/she gets in the job. People want to know that their efforts are making a difference and aren't going unnoticed by high-ups. Rewarding and recognizing employees creates stronger relationships, which in turn spurs motivation. And finally, as mentioned previously, rewarding employees for their work motivates them to stay.

- The organization does impact on salary as the p value is less than 0.05, so they are dependent on each other. The bigger the organization the more the salary is expected from the company.

- The management decisions of a company put impact on the job stress as the p value is less than 0.05 and the decisions taken by the management is for the betterment of the company to increase their profits and business which will put stress on the employees. The stress experienced by workers can cause short- and long-term problems that can eventually cause a business to go under.

- Substantial progress has been made to narrow the pay gap. Women’s wages are now significantly closer to men’s, but in recent years, that progress has stalled. Greater gender equity and increasing female economic participation are associated with higher growth, more favourable development outcomes, and lower income inequality.

- An organization name can impact both your present and future career in several ways. It can indicate how much experience u gained and how much authority and how much skills you must lead a team.

- Workshops help the employees in gaining the right set of skills and abilities to perform better and thus improve their performance.

- When employees are given the freedom to work or solve a problem on their own, they are likely to think of a solution that is unique and involves their own thought process which directly impact on an organization in a positive way.

- Age doesn't matter. In a professional world, age doesn't matter if the person is equipped with the right skill set, with good experience and better at work.

- The age factor of an employee is particularly not considered while providing day care benefits. Hence, the two factors are independent of each other. As, some of the employees with same age are provided with different benefits.

- Not every employee with high annual pays proud of the work that they do at the organization. Some employees are content with their work but mostly employees remained working at the company just to meet the financial obligations.

- The work environment is safe in the organization except for one or two cases related to age discrimination and gender discrimination. All the other employees did not face any kind of discrimination at the organization.