DATA PRODUCT ON REMOTE WORK AND MENTAL HEALTH

ANALYZING FACTORS INFLUENCING EMPLOYEE SATISFACTION WITH REMOTE WORK

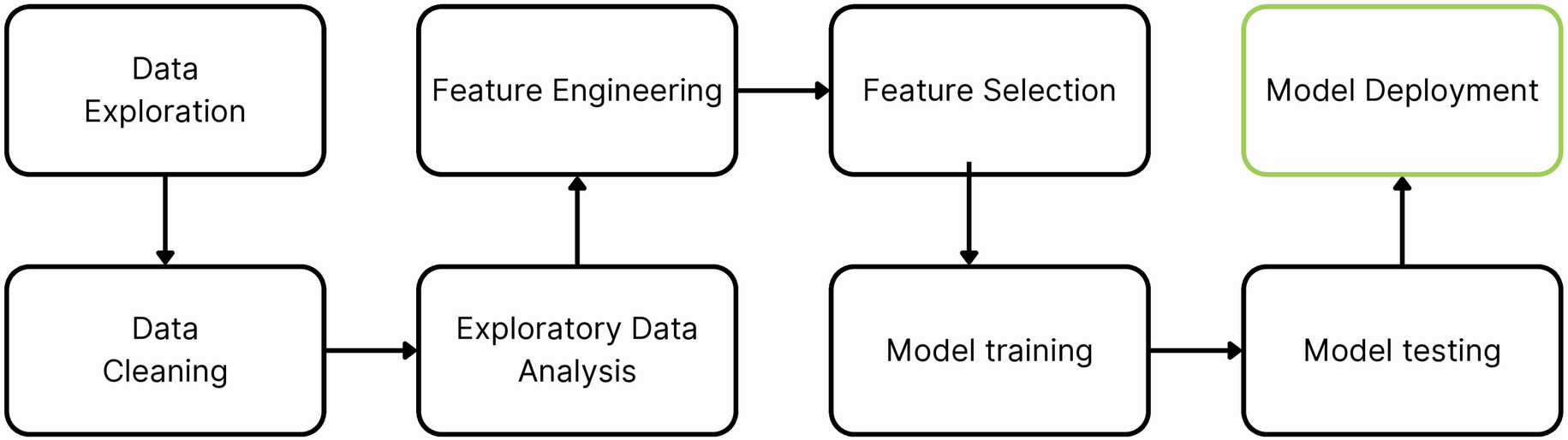
OBJECTIVE OF THE ASSIGNMENT

- Build a data product around the Kaggle dataset on remote work and mental health.

- Address key questions regarding factors affecting employee satisfaction.
    - What factors influence employee satisfaction in remote work?
    - How to predict employee satisfaction based on identified factors?

Dataset Source: Kaggle: Remote Work and Mental Health

Protocol


Data Exploration
Data Cleaning
Exploratory Data Analysis
Feature Engineering
Feature Selection
Model training
Model testing
Model Deployment

Data Info

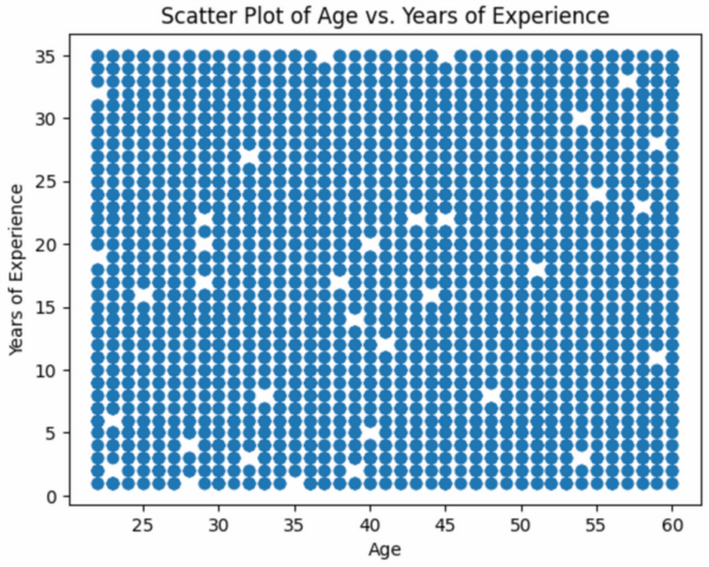
Column	Non null	Datatype	Unique
Employee_ID	5000	String	5000
Age	5000	Integer	39
Gender	5000	String	4
Job_Role	5000	String	7
Industry	5000	String	7
Years_of_Experience	5000	Integer	35
Work_Location	5000	String	3
Hours_Worked_Per_Week	5000	integer	41
Number_of_Virtual_Meetings	5000	integer	16
Work_Life_Balance_Rating	5000	integer	5
Stress_Level	5000	String	3
Mental_Health_Condition	3804	String	3
Access_to_Mental_Health_Resources	5000	String	2
Productivity_Change	5000	String	3
Social_Isolation_Rating	5000	integer	5
Satisfaction_with_Remote_Work	5000	String	3
Company_Support_for_Remote_Work	5000	integer	5
Physical_Activity	3371	String	2
Sleep_Quality	5000	String	3
Region	5000	String	6

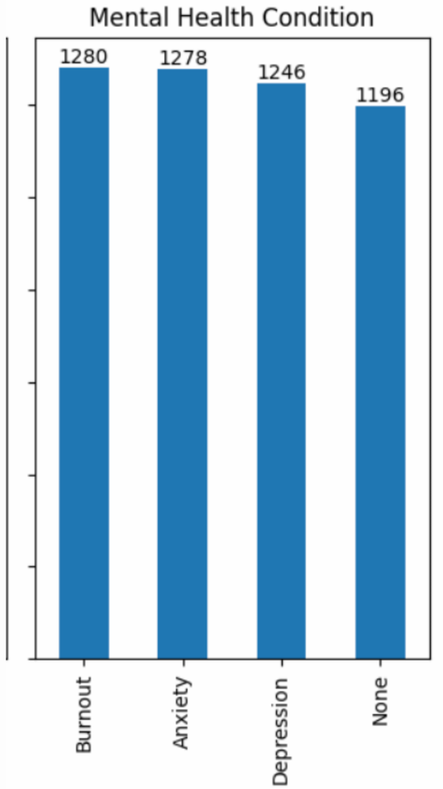
Key Points

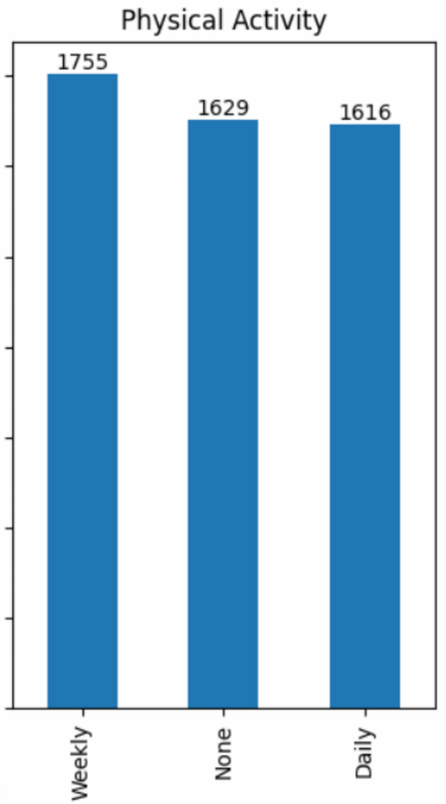
1. Employee_ID: all data is unique don't add anything to the model
2. Most of the columns are not null
3. Physical_Activity and Mental_Health_Condition are having null values
4. Age ,Years_of_Experience ,Hours_Worked_Per_Week ,Number_of_Virtual_Meetings ,Work_Life_Balance_Rating ,Social_Isolation_Rating ,Company_Support_for_Remote_Work are integer, but we can't say which are ordinal
5. Gender ,Job_Role ,Industry ,Work_Location ,Stress_Level ,Mental_Health_Condition ,Access_to_Mental_Health_Resources ,Productivity_Change ,Satisfaction_with_Remote_Work ,Physical_Activity ,Sleep_Quality ,Region are strings so they need to be encoded.

Data Cleaning

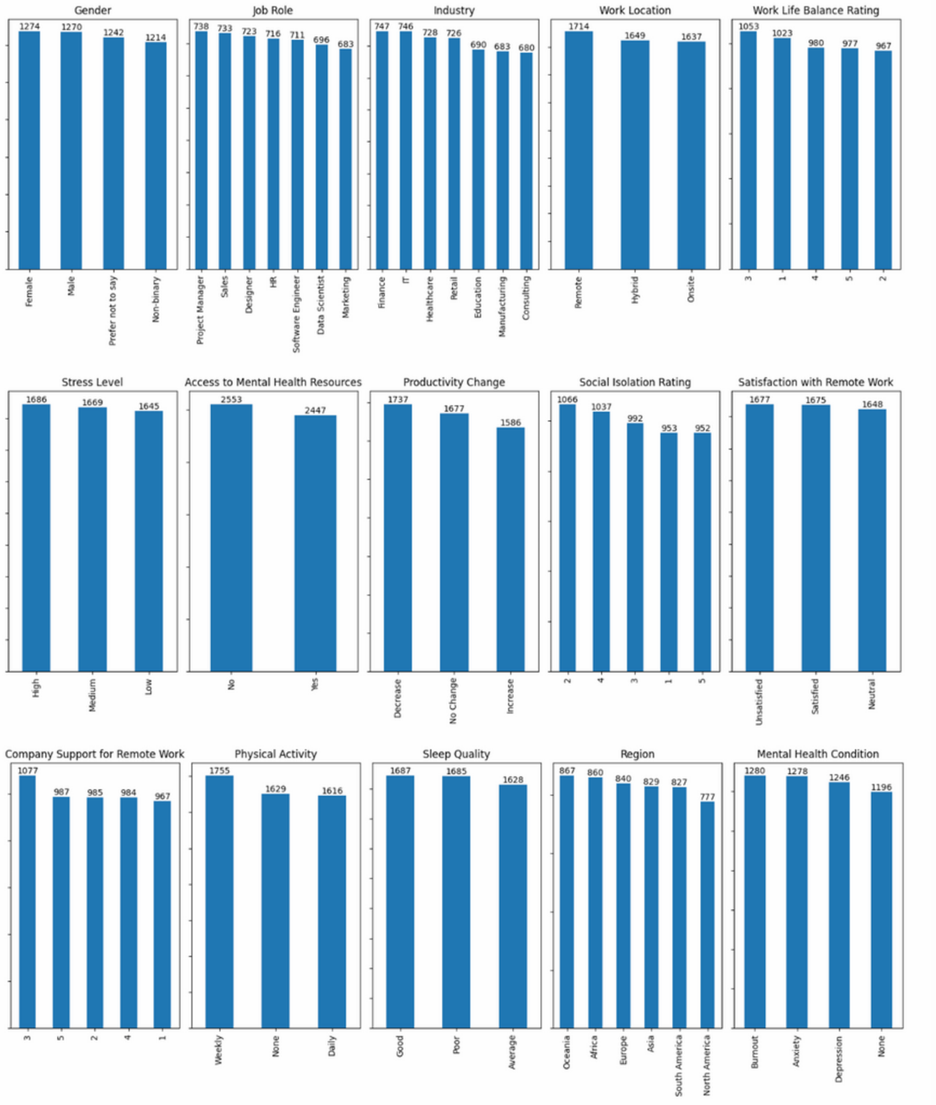
1. On checking data I have found the data is clean according to the syntax but logically it is having issue when we check year of experience and Age which shows we have 25 age person with 35 year of experience which don't make sense. For cleaning this we have removed anyone who is having more than 18 years of between Age and Experience considering an individual can start working from age 18.
2. Physical_Activity and Mental_Health_Condition are having None in them which make sense and we considered it as employees are not active and not feeling any tirednes, or burnout that's why we got None, so we replaced Nan as None





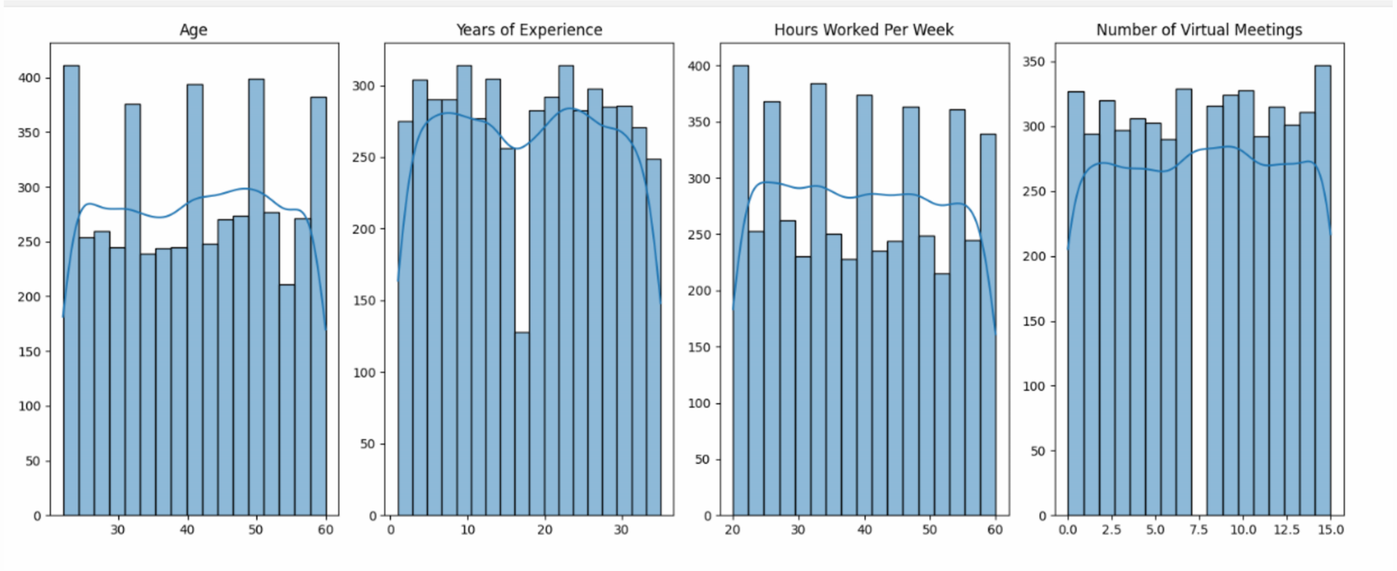


Exploratory Data Analysis (EDA)



1- On a quick glimpse we can see that data is very balance in both numerical and string columns, so we don't have to take care of class imbalaning .

2- This behavior is same when we checked interaction between different features.



String Feature check

Gender	Job_Role	Industry	Work_Location	Stress_Level	Mental_Health_Condition	Access_to_Mental_Health_Resources	Productivity_Change	Satisfaction_with_Remote_Work	Physical_Activity	Sleep_Quality	Region
Non-binary	HR	Healthcare	Hybrid	Medium	Depression	No	Decrease	Unsatisfied	Weekly	Good	Europe
Female	Data Scientist	IT	Remote	High	Anxiety	Yes	Increase	Satisfied	None	Poor	Asia
Male	Software Engineer	Education	Onsite	Low	None		No Change	Neutral	Daily	Average	North Americ
Prefer not to say	Sales	Finance			Burnout						South Americ
	Marketing	Consulting									Oceania
	Designer	Manufacturing									Africa
	Project Manager	Retail									

1. If we see columns like Stress_Level, Productivity_Change,Satisfaction_with_Remote_Work and Sleep_Quality, this are the values can be scaled and this we can scale with Ordinal Encoder
2. Whereas as columns like Gender, we can’t say Female is 3 and Male is 2, so this columns we will encode with one hot encoder

Feature Categories

CATEGORICAL

Gender
Job_Role
Industry
Work_Location
Mental_Health_Condition
Access_to_Mental_Health_Resources
Physical_Activity
Region

ORDINAL

Stress_Level
Productivity_Change
Satisfaction_with_Remote_Work
Sleep_Quality

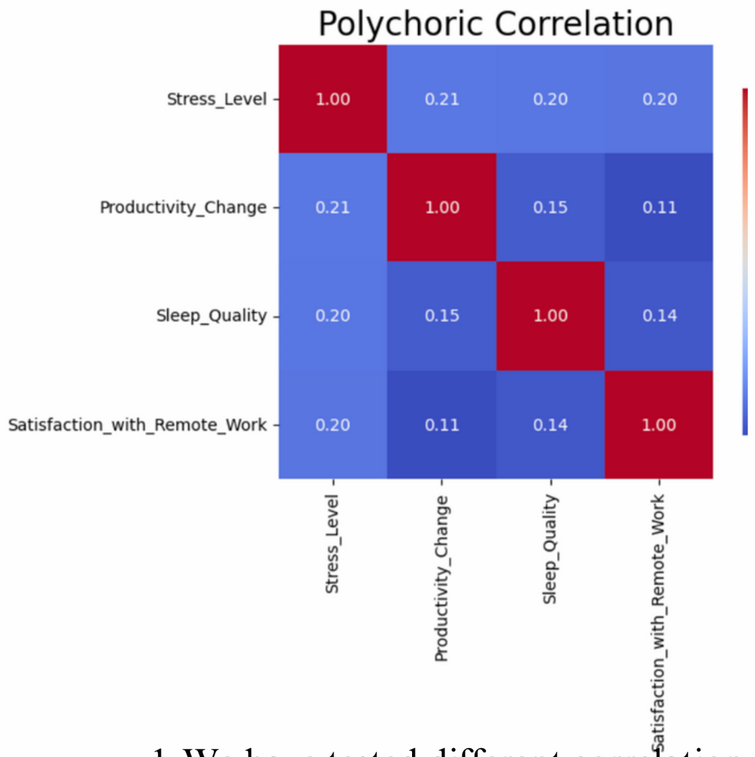
NUMERICAL

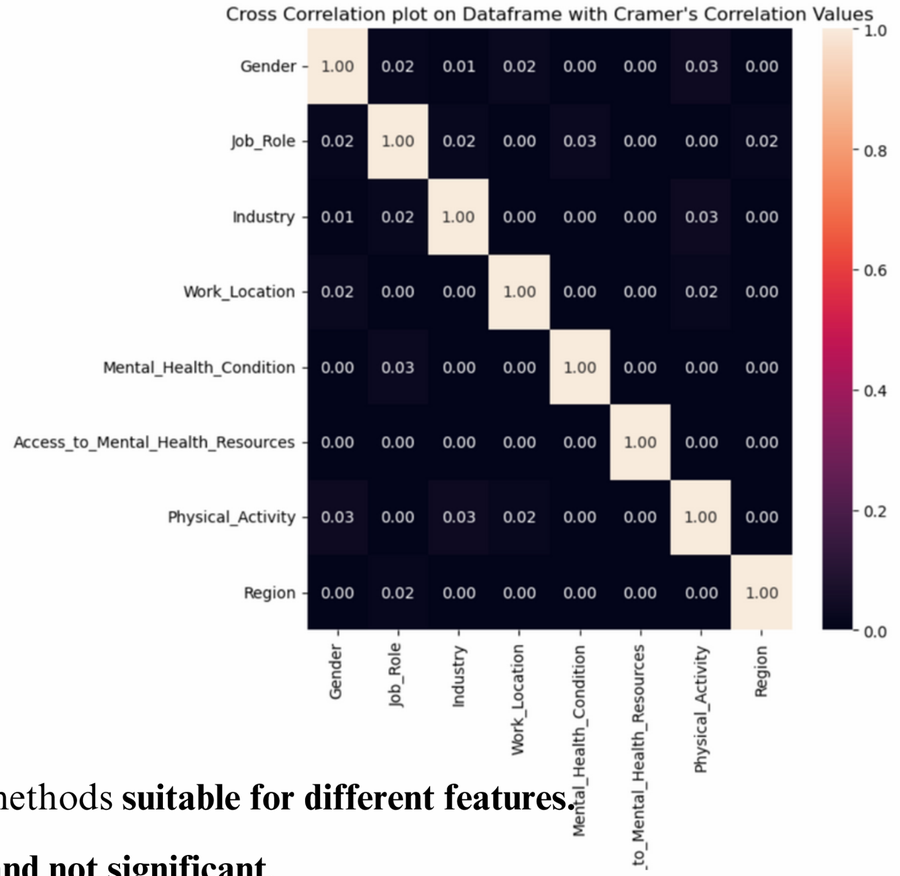
Age
Years_of_Experience
Hours_Worked_Per_Week
Number_of_Virtual_Meetings

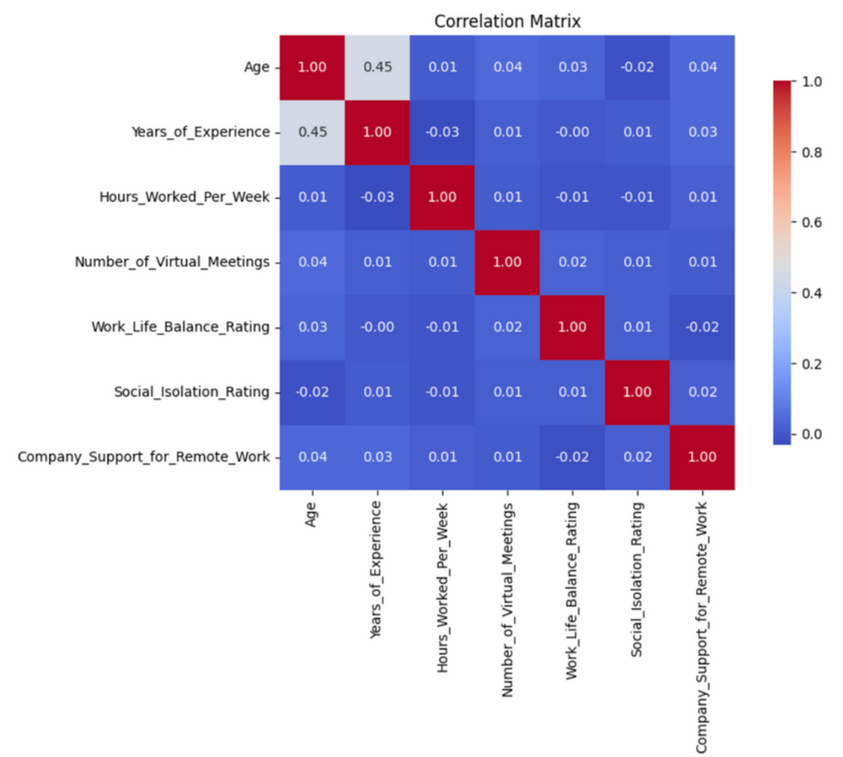
Target

Satisfaction_with_Remote_Work

Correlation







1. We have tested different correlation methods **suitable for different features.**
2. We identified **correlation is very less and not significant**.
3. So for identifying important feature, used random forest after encoding and scaling to idenify the importance of each feature, which is itself not significant.
4. We scaled every ordinal values except the target as, we decided to use classifier to find which class they belong to and also, **we need to see that in this classification, we need to see, by how much we are wrong**. For example. if a employee is satisfied with the the remote work, but if we predict unsatisfied it will completely wrong but identifying as Neutral can be better result.

Feature Imortance

Feature	Importance
Hours_Worked_Per_Week	0.089146
Years_of_Experience	0.087274
Age	0.085568
Number_of_Virtual_Meetings	0.075396
Company_Support_for_Remote_Work	0.049098
Social_Isolation_Rating	0.047085
Work_Life_Balance_Rating	0.047075
Stress_Level	0.033506
Productivity_Change	0.03306
Sleep_Quality	0.033035

Key Points

1. We can see that feature importance is not much but we have chosen top 10 features and trained it using this.
2. After this we have used Weight and bias Sweep method to run different models like **RandomForestClassifier, XGBoostClassifier and SVC** for classification and identified the best result which we recevied is little above 33% which supports, that the data is not proper for prediction.
3. Feature_engineering and Training model scripts will be found in GITHUB

Conclusion

1. After overall analysis it can be said that data is balanced artificially and without any dependency
2. With the random data, it is hard to reterive any valuable insights