INSURANCE PREMIUM AND RISK ANALYSIS

Data Analytics Project Report

1.INTRODUCTION:

The insurance industry needs to understand customer risk and medical costs to improve pricing and profitability. This project focuses on analyzing insurance customer data to identify key factors that affect insurance charges and risk levels. The analysis was performed using **Python, SQL, and Power BI**.

2. PROBLEM STATEMENT:

The main objectives of this project are:

- To understand what factors, increase insurance charges
- To identify high-risk and low-risk customers
- To analyze the impact of age, BMI, smoking, children, gender, and region
- To support better decision-making for insurance premium and risk management

3.DATASET DESCRIPTION:

The dataset contains information about insurance customers, including:

- **age** – customer age
- **gender** – male or female
- **bmi** – body mass index
- **children** – number of dependents
- **smoker** – smoking status (yes/no)
- **region** – residential area
- **charges** – medical insurance charges

Total records: -1,337 customers

4. DATA CLEANING:

Data cleaning was performed using Python to ensure accurate analysis:

- Checked and removed duplicate records
- Handled missing or invalid values
- Verified correct data types
- Identified extreme values (outliers) in insurance charges
- Ensured consistency in categorical values (smoker, region, gender)

5. EXPLORATORY DATA ANALYSIS (EDA):

EDA was used to understand data patterns and distributions:

- Analyzed average, minimum, and maximum insurance charges
- Studied the distribution of customers by risk level
- Compared charges across age groups, BMI categories, and smoking status
- Identified customers with very high insurance costs

Key Observations:

Insurance charges are not evenly distributed and are strongly influenced by lifestyle factors.

6. FEATURE ENGINEERING:

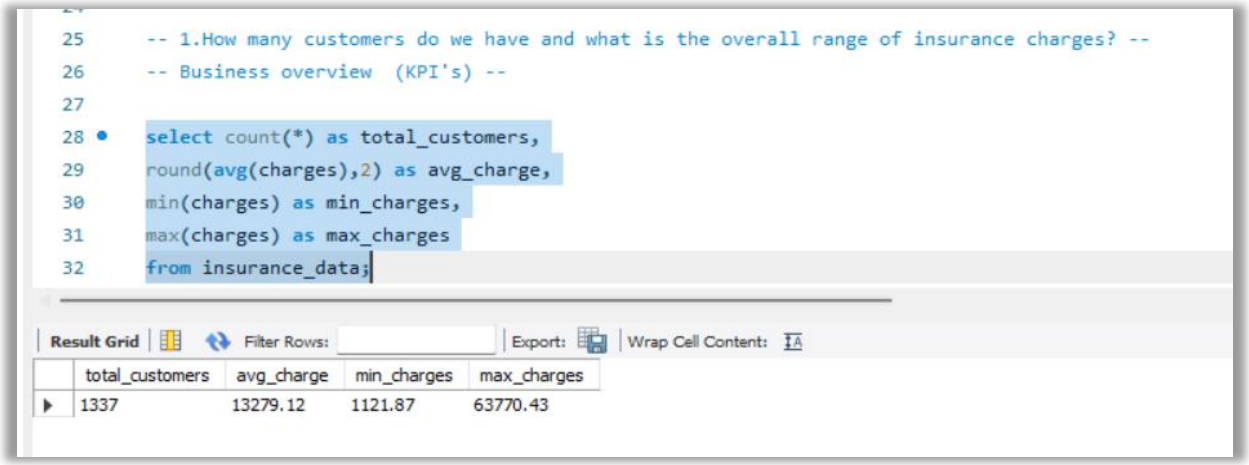
New features were created to improve analysis:

- **BMI Category**: Underweight, Normal, Overweight, Obese
- **Age Group**: Below 25, 25–40, 41–60, 60+
- **Risk Level**: High Risk and Low Risk (based on charges and health indicators)

These features helped in better segmentation and visualization.

7. INSIGHTS:

I. Overall Business Overview (KPI Summary)



```
25    -- 1.How many customers do we have and what is the overall range of insurance charges? --
26    -- Business overview  (KPI's) --
27
28    select count(*) as total_customers,
29    round(avg(charges),2) as avg_charge,
30    min(charges) as min_charges,
31    max(charges) as max_charges
32    from insurance_data;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

total_customers	avg_charge	min_charges	max_charges
1337	13279.12	1121.87	63770.43

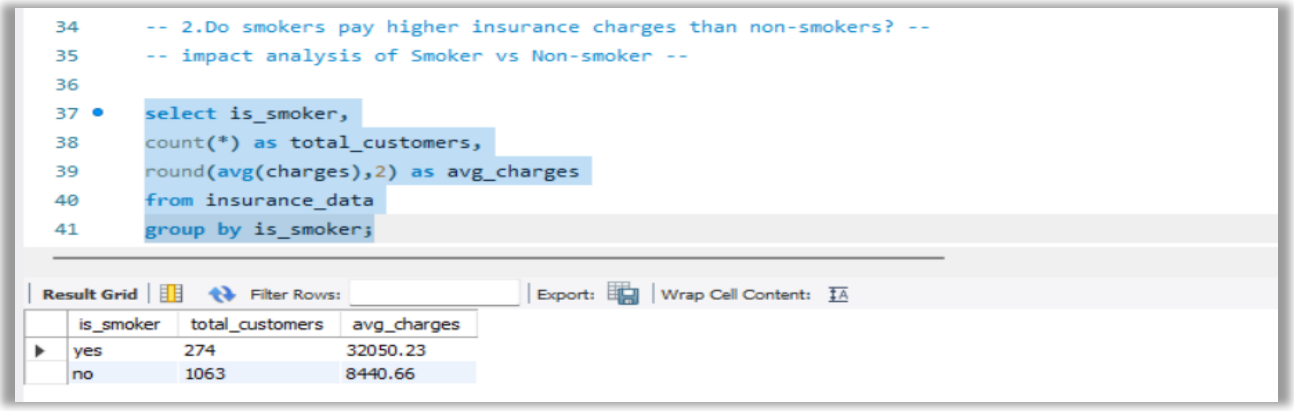
Result:

- Total customers: **1,337**
- Avg charges: **₹13,279**
- Min charges: **₹1,122**
- Max charges: **₹63,770**

Insight:
Insurance charges vary widely, indicating the presence of **low-cost and high-cost customer segments**.

Action/Suggestion:
Create **risk-based customer segments** and focus on monitoring **high-cost customers**.

II. Impact of Smoking on Insurance Charges



```
34    -- 2.Do smokers pay higher insurance charges than non-smokers? --
35    -- impact analysis of Smoker vs Non-smoker --
36
37    select is_smoker,
38    count(*) as total_customers,
39    round(avg(charges),2) as avg_charges
40    from insurance_data
41    group by is_smoker;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

is_smoker	total_customers	avg_charges
yes	274	32050.23
no	1063	8440.66

Result:

- Smokers: **274 customers | Avg charges = ₹32,050**
- Non-smokers: **1,063 customers | Avg charges = ₹8,441**

Insight:
Smokers have **almost 4 times higher insurance charges** than non-smokers. Smoking is a **major cost-driving risk factor**.

Action/Suggestion:
Introduce **higher premium slabs for smokers** and launch **preventive health programs** to reduce long-term risk.

III. Gender-wise Insurance Charges

```
43    -- 3.Is there a big difference in insurance charges between male and female customers? --
44    -- Gender wise average charge analysis --
45
46    select gender,
47    count(*) as total_customers,
48    round(avg(charges),2) as avg_charges
49    from insurance_data
50    group by gender;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

gender	total_customers	avg_charges
female	662	12569.58
male	675	13975.00

Result:

- Female: **662 customers | Avg charges = ₹12,569**
- Male: **675 customers | Avg charges = ₹13,975**

Insight:
Male customers have **slightly higher average insurance** charges than female customers. However, the difference is **not very large**, which indicates that **gender alone is not a strong cost driver**.

Action/Suggestion:
Do not design pricing mainly on gender. Instead, focus more on **health and lifestyle factors** such as smoking status, BMI, and age.

iv. Region-wise Business Performance

```
52    -- 4. Which region generates the highest insurance revenue? --
53    -- Region wise business performance analysis --
54
55    select region,
56    count(*) as total_Customers,
57    round(avg(charges),2) as avg_charge,
58    round(sum(charges),2) as total_revenue
59    from insurance_data
60    group by region
61    order by total_revenue desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

region	total_Customers	avg_charge	total_revenue
southeast	364	14735.41	5363689.80
northeast	324	13406.38	4343668.64
northwest	324	12450.84	4034072.37
southwest	325	12346.94	4012754.82

Result:

- Southeast: - 364 customers | Avg ₹14,735 **(Highest revenue)**
- Northeast: - 324 customers | Avg ₹13,406
- Northwest: - 324 customers | Avg ₹12,451
- Southwest: - 325 customers | Avg ₹12,347 **(Lowest revenue)**

Insight:

The **Southeast region generates the highest insurance revenue** and also shows the **highest average charges.** Other regions have similar customer counts, but lower average charges and revenue.

Action / Suggestion:

- Focus **marketing and premium products** in the Southeast region.
- Analyze why other regions have lower charges and design **region-specific pricing and health programs**.

v: Impact of Number of Children on Insurance Charges

```
63      -- 5.Does the number of children affect insurance charges? --
64      -- Children count impact on insurance premium --
65
66  •   select children,
67      count(*) as total_customers,
68      round(avg(charges),2) as avg_charge
69      from insurance_data
70      group by children
71      order by avg_charge desc ;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

children	total_customers	avg_charge
3	157	15355.32
2	240	15073.56
4	25	13850.66
1	324	12731.17
0	573	12384.70
5	18	8786.04

Result:

- 3 children → Avg charges ₹15,355 **(highest, sample value)**
- 2 children → Avg charges ₹15,074
- 4 children → Avg charges ₹13,851
- 1 child → Avg charges ₹12,731
- 0 children → Avg charges ₹12,385
- 5 children → Avg charges ₹8,786 **(lowest, small sample)**

Insight:

Customers with **2–3 children show higher average insurance charges.** However, the pattern is **not strictly linear**, which means children count alone is **not a strong cost driver**, but it slightly influences premium.

Action/Suggestion:

Consider children count as a **supporting factor** in pricing, but rely more on **age, smoking, and BMI** for primary risk assessment.

vi. Impact of BMI Category on Insurance Charges

```
-- 6.Do customers with higher BMI pay more insurance charges? --
-- category wise BMI analysis --


select bmi_category,
count(*) as total_people,
round(avg(charges),2) as avg_charge

from insurance_data
group by bmi_category
order by avg_charge desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

bmi_category	total_people	avg_charge
obese	706	15572.04
overweight	386	10987.51
Normal	225	10409.34
Underweight	20	8852.20

Result:

- Obese → 706 people | Avg charges ₹15,572 **(highest)**
- Overweight → 386 people | Avg charges ₹10,988
- Normal → 225 people | Avg charges ₹10,409
- Underweight → 20 people | Avg charges ₹8,852 **(lowest)**

Insight:

Insurance charges **increase significantly with higher BMI levels**. Obese customers form the **largest group** and also generate the **highest average charges**, making BMI a **strong health-based cost driver**.

Action / Suggestion:

- Include BMI as a key risk factor in premium calculation.
- Launch wellness and weight-management programs to reduce long-term claim costs.

vii. High-Risk vs Low-Risk Customer Analysis

```
-- 7.How much higher are the insurance charges for high-risk customers compared to low-risk customers? --
-- High risk vs low risk customer analysis --

select risk_level,
count(*) as total_customers,
round(avg(charges),2)as avg_charge
from insurance_data
group by risk_level
order by avg_charge desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

risk_level	total_customers	avg_charge
High risk	144	41692.81
low risk	1193	9849.47

Result:

- High Risk (1): 144 customers | Avg charges **₹41,693**
- Low Risk (0): 1,193 customers | Avg charges **₹9,849**

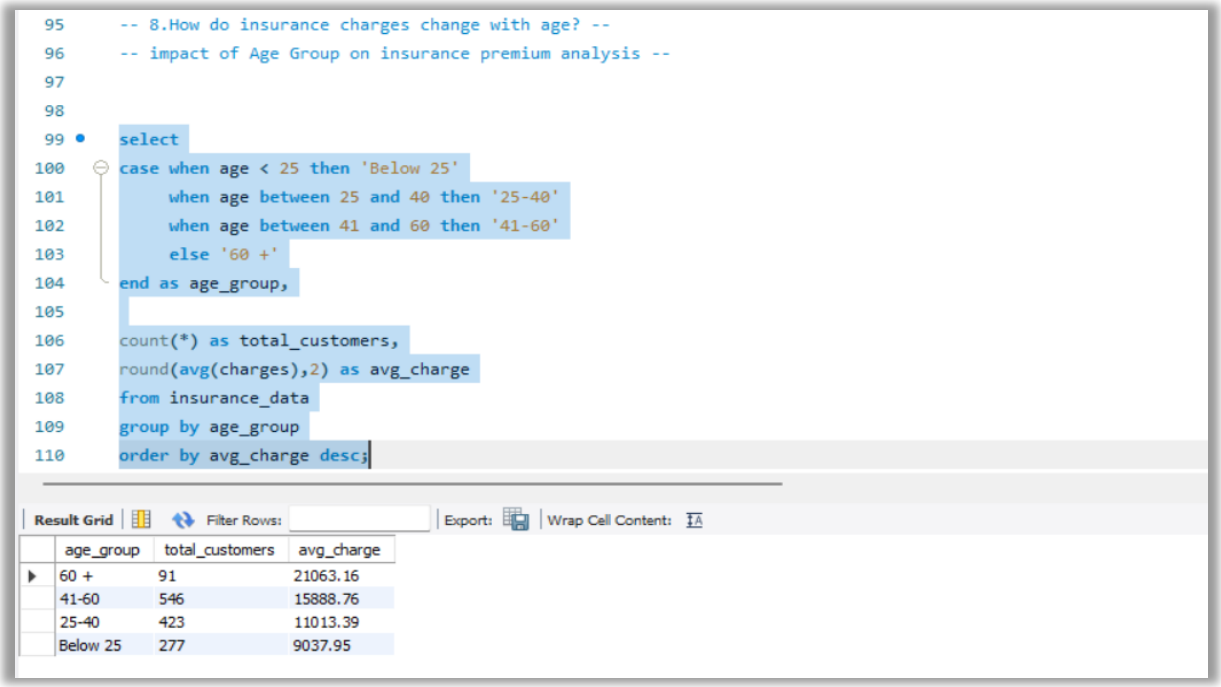
Insight:

High-risk customers are very few in number, but their average insurance cost is more than 4 times higher than low-risk customers. This shows that a small high-risk segment contributes disproportionately to total claims.

Action / Suggestion:

- Design separates high-risk insurance plans with higher premiums.
- Invest in early screening and preventive care programs to control high-risk costs.

viii. Impact of Age Group on Insurance Charges



```
95   -- 8.How do insurance charges change with age? --
96   -- impact of Age Group on insurance premium analysis --
97
98
99   select
100  case when age < 25 then 'Below 25'
101       when age between 25 and 40 then '25-40'
102       when age between 41 and 60 then '41-60'
103       else '60 +'
104  end as age_group,
105
106  count(*) as total_customers,
107  round(avg(charges),2) as avg_charge
108  from insurance_data
109  group by age_group
110  order by avg_charge desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

age_group	total_customers	avg_charge
60 +	91	21063.16
41-60	546	15888.76
25-40	423	11013.39
Below 25	277	9037.95

Result:

- 60+ = 91 customers **|** Avg charges **₹21,063 (highest)**
- 41–60 = 546 customers | Avg charges **₹15,889**
- 25–40 = 423 customers | Avg charges **₹11,013**
- Below 25 = 277 customers | Avg charges **₹9,038 (lowest)**

Insight:

Insurance charges increase clearly with age. Customers above 60 have the highest average medical cost, making age a strong and consistent risk factor.

Action / Suggestion:

- Apply **age-based premium slabs**.
- Provide **senior-focused health plans and preventive care programs** to manage long-term costs.

ix. Top 10 Highest Paying Customers Pattern

```sql
-- 9. Who are the top 10 customers with the highest insurance charges? --
--  Top 10 highest paying customers --

SELECT age, gender, is_smoker, bmi, children, charges
FROM insurance_data
ORDER BY charges DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

age	gender	is_smoker	bmi	children	charges
54	female	yes	47.41	0	63770.43
45	male	yes	30.36	0	62592.87
52	male	yes	34.49	3	60021.40
31	female	yes	38.10	1	58571.07
33	female	yes	35.53	0	55135.40
60	male	yes	32.80	0	52590.83
28	male	yes	36.40	1	51194.56
64	male	yes	36.96	2	49577.66
59	male	yes	41.14	1	48970.25
44	female	yes	38.06	0	48885.14

Result:

Top 10 customers have charges between **₹48,885 and ₹63,770**. All top 10 customers are **smokers**, and most of them have **high BMI (overweight/obese)** and belong to **middle-aged or senior groups**.

Insight:

The highest-paying customers share **common risk characteristics** smoking habit, high BMI, and higher age. This clearly shows that **lifestyle and health factors dominate high insurance costs**.

Action / Suggestion:

- Build a **high-risk customer profile** for early identification.
- Design **special high-premium plans** and **health improvement programs** for such customers.

x. Region-wise Top Costly Smokers

```sql
-- 10. Who is the highest-cost smoker in each region? --
-- region wise top costly smoker persons --

select * from
(select region,age,gender,charges,
rank() over(partition by region order by charges desc )as rnk
from insurance_data
where is_smoker = 'yes'
) t

where rnk = 1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

region	age	gender	charges	rnk
northeast	31	female	58571.07	1
northwest	52	male	60021.40	1
southeast	54	female	63770.43	1
southwest	60	male	52590.83	1

Result:

For each region, the highest-cost smoker is:

- Northeast = **₹58,571**
- Northwest = **₹60,021**
- Southeast = **₹63,770 (highest overall)**
- Southwest = **₹52,591**

Insight:

Every region has **at least one extremely high-cost smoker**, and the **Southeast region shows the highest individual medical cost**. This indicates that **smokers are the most financially risky group across all regions**.

Action / Suggestion:

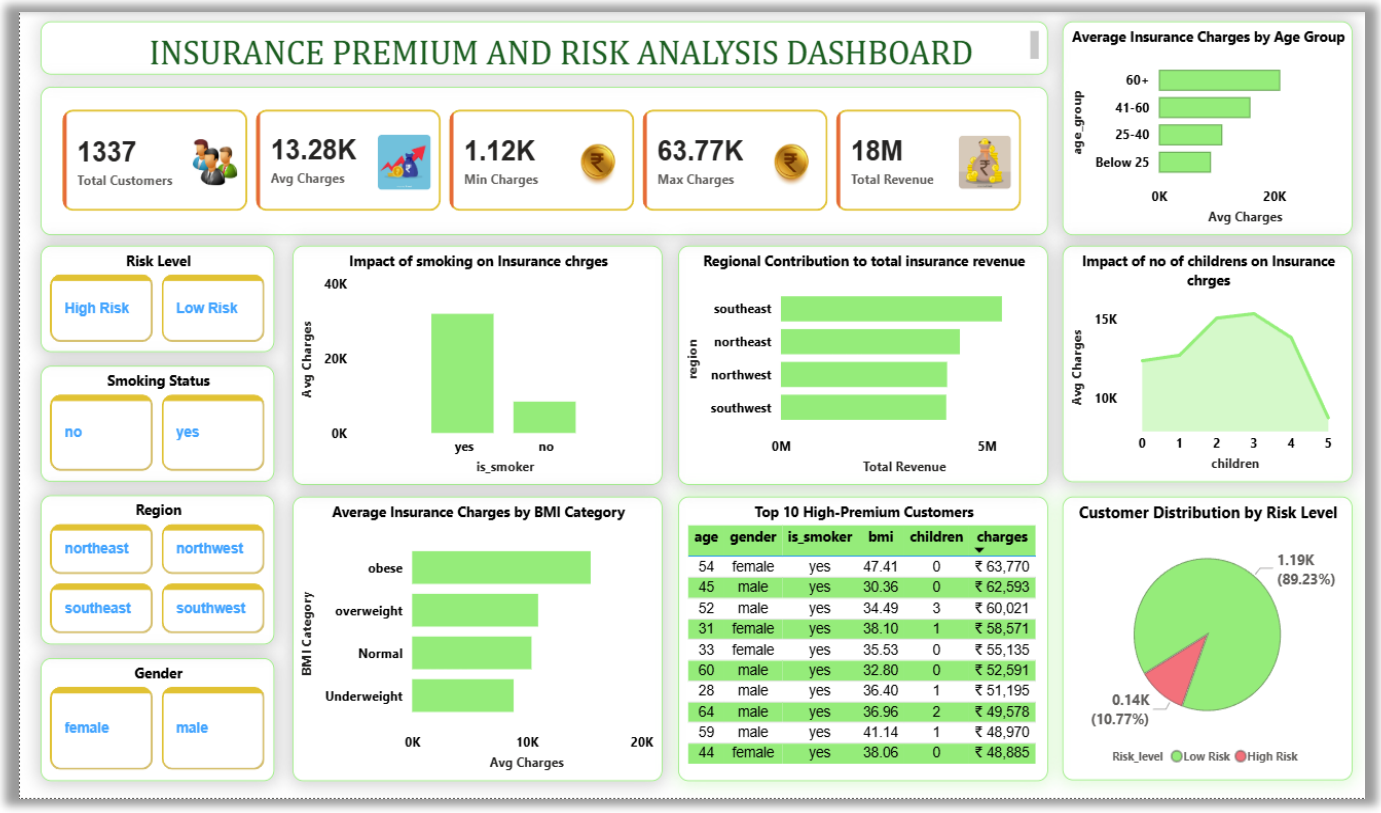
- Identify and **continuously monitor region-wise high-risk smokers**.
- Design **region-specific high-risk insurance plans** and **targeted preventive health programs**.

8.KEY INSIGHTS FROM THE ANALYSIS:

- **Smokers have significantly higher insurance charges than non-smokers**
- **Obese customers show the highest average charges**
- Insurance charges increase with **age**
- Customers with **2–3 children** show slightly higher average charges
- **High-risk customers form a small percentage but contribute a large portion of total revenue**
- Southeast region contributes the highest insurance revenue

9. VISUALIZATION:

A Power BI dashboard was created to present insights clearly:



- KPI cards showing total customers, average charges, minimum and maximum charges, and total revenue
- Charts for:
    - Charges by BMI category
    - Impact of smoking on charges
    - Charges by age group
    - Revenue contribution by region
- Interactive slicers for region, smoking status, gender, and risk level
- Table showing **Top 10 High-Premium Customers**

The dashboard allows dynamic filtering and easy interpretation.

10. BUSINESS RECOMMENDATIONS:

Based on the analysis:

- Introduce **higher premiums or special policies for smokers**
- Promote **preventive health programs** for obese and high-risk customers
- Offer **discounts or incentives** to low-risk customers
- Use risk segmentation to design **customized insurance plans**
- Focus on high-revenue regions for targeted marketing

11. CONCLUSION:

This project successfully analyzed insurance customer data to identify cost drivers and risk factors. By combining Python, SQL, and Power BI, meaningful insights were generated that can help insurance companies improve pricing strategies, manage risk, and increase profitability.