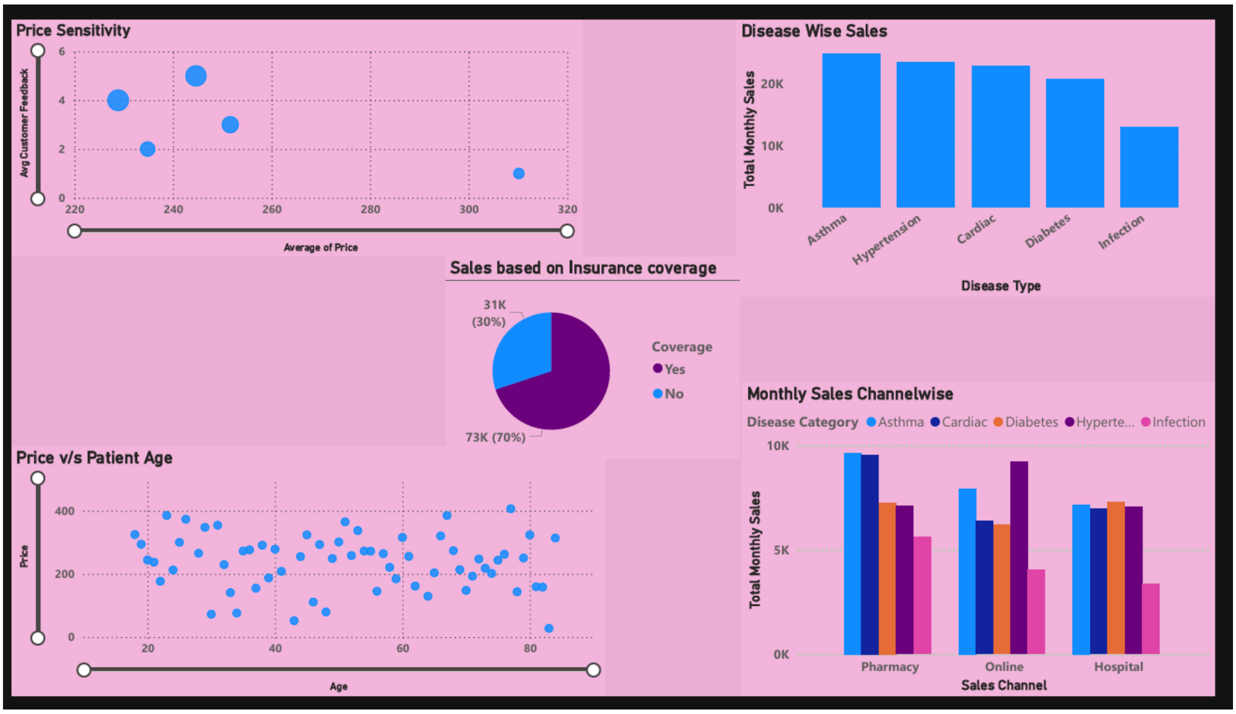
Comprehensive Report on Healthcare Data Dashboard and Insights



1. Business Problem

The healthcare industry has seen a massive transformation, particularly in the pharmaceutical sector, driven by advancements in technology, evolving consumer demands, and increasing focus on cost-effective, personalized care. One of the challenges pharmaceutical companies face is accurately pricing their products to meet customer expectations while maintaining profitability. Additionally, understanding how sales are influenced by insurance coverage, disease categories, and sales channels is essential to optimize business strategies.

This project is designed to address this business problem by providing insights that can directly impact pricing strategies, sales channels, and product offerings. By understanding customer purchasing behavior, healthcare providers can better align their products with market demand. The synthetic dataset was created to simulate real-world conditions, allowing for comprehensive analysis without any privacy or data security concerns. This is important in a sector where sensitive information must be handled with care.

The synthetic nature of the dataset also ensures that we have controlled variability, making it easier to test hypotheses and draw meaningful conclusions. The insights gained will inform pharmaceutical companies on where to focus their resources, what products to promote, and how to improve customer satisfaction through pricing adjustments.

2. Data Requirement

The data required for this study needed to reflect a wide range of variables that influence the pharmaceutical sales ecosystem. These included customer demographics (age, insurance status), product attributes (price, disease category), and sales performance indicators (monthly sales, feedback). This wide-ranging data allows us to comprehensively analyze factors that influence purchasing behavior.

The inclusion of customer feedback was crucial, as it provides qualitative insight into how customers perceive pricing and product quality. Feedback is often one of the strongest predictors of future purchase behavior, so understanding customer sentiment helps in refining products and marketing strategies. Additionally, by including sales data categorized by disease type and sales channels, we can understand which segments of the market are driving revenue, which can inform targeted marketing campaigns.

The choice of these specific variables was driven by their relevance to the pharmaceutical industry, where purchasing decisions are influenced not only by price but also by disease-related needs, insurance coverage, and available sales channels.

3. Data Collection and Data Understanding

The dataset was synthetically generated using Generative AI to ensure a realistic yet controlled environment for analysis. The synthetic nature of the data was chosen to avoid any privacy concerns that come with using real patient or sales data. Furthermore, synthetic data offers the flexibility to simulate a wide range of scenarios, including edge cases, which might not be readily available in real-world datasets.

In generating the dataset, various factors such as disease prevalence, price ranges for pharmaceutical products, and common customer demographics were taken into account to ensure that the dataset accurately reflected typical market conditions. The generation of synthetic data also allowed for easier data manipulation and ensured that all variables were structured in a way that could be easily analyzed and visualized.

The decision to use a synthetic dataset was further justified by the need for an unbiased, controlled environment where we could test various assumptions without being constrained by the limitations of real-world data. This allows for a more flexible and creative exploration of potential business solutions.

4. Data Validation

Data validation is crucial to ensuring that the results of any analysis are both accurate and reliable. Given that the dataset was synthetic, validating the data was essential to confirming that the simulated relationships between variables (such as price and customer feedback) were logically consistent and reflective of real-world patterns.

To achieve this, we performed a variety of checks. For numerical fields like age and sales figures, we used range validation to ensure that no values were out of expected bounds. For categorical data, such as disease categories or sales channels, consistency checks were performed to verify that all entries adhered to a pre-defined set of valid values. This validation process ensured that any analyses performed on the dataset would be grounded in valid data, giving us confidence in the insights generated.

The choice to perform these checks was rooted in the need to maintain data integrity and avoid introducing errors or biases into the analysis. By ensuring that the data was accurate, we could rely on the resulting insights to drive effective business decisions.

---

5. Data Cleaning

Data cleaning is an essential step in preparing any dataset for analysis. Since the data was synthetic, it was created with the assumption that some inconsistencies or errors could arise, which needed to be addressed before meaningful analysis could take place.

For example, missing values in the dataset were handled by imputing them with the mean or mode, depending on whether the variable was continuous or categorical. This approach ensures that the dataset remains complete without introducing bias into the analysis. For categorical variables, like disease categories or insurance coverage, any inconsistencies in spelling or formatting were corrected to ensure uniformity across records.

The reason for choosing this cleaning method was that imputing values for missing data avoids the issue of incomplete datasets, which could skew the results. Removing outliers and addressing inconsistencies also ensured that the dataset remained representative of typical market conditions, making the subsequent analysis more accurate.

---

6. Tools Used

The tools chosen for this project were Microsoft Excel and Power BI, which were selected based on their versatility and ease of use for data analysis and visualization.

Microsoft Excel was used during the data cleaning phase. Its robust set of functions allowed us to perform calculations, handle missing data, and format the dataset in preparation for visualization. Excel is also widely accessible and familiar to many users, making it an ideal tool for preprocessing and organizing the data.

Power BI was chosen for the visualization phase due to its dynamic and interactive capabilities. Unlike static tools, Power BI enables the creation of dashboards that allow users to interact with the data in real-time. The use of filters and slicers allows stakeholders to drill down into specific segments, such as disease categories or age groups, providing a more tailored and flexible view of the data.

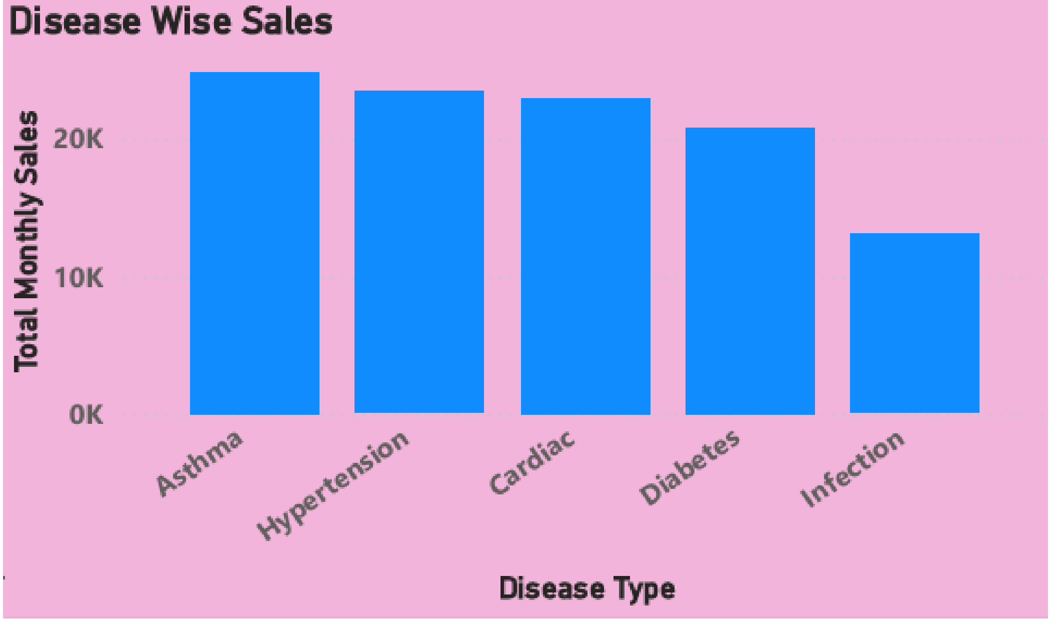
These tools were chosen because they enable a seamless workflow from data preparation to presentation. They also offer powerful features that ensure the results are both insightful and easily interpretable by stakeholders.

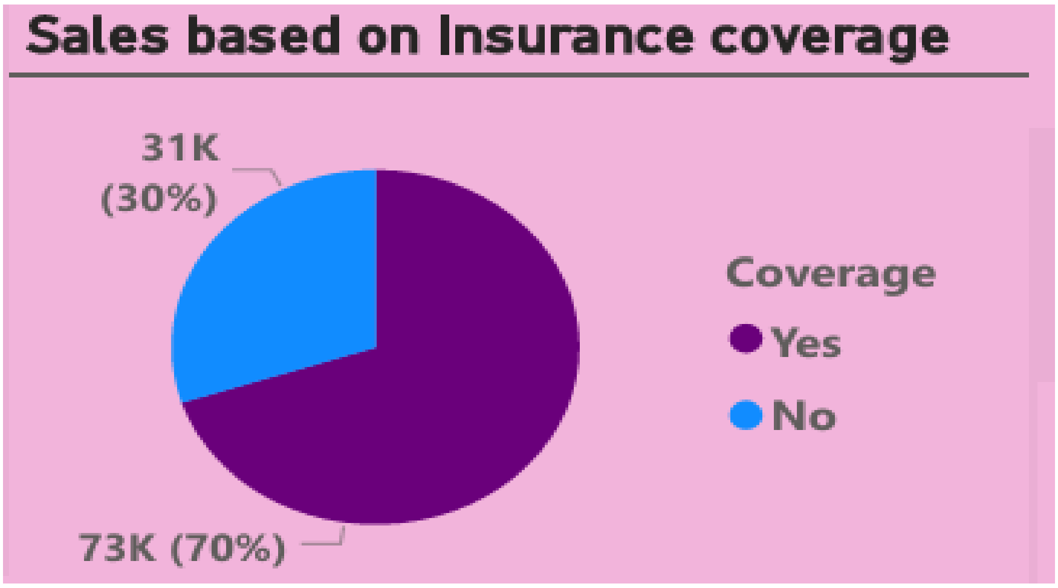
7. Graphs and Charts

Various graphs and charts were created to explore the relationships between different variables. The decision to use these visualizations was driven by the need to present complex data in an easily digestible format, allowing stakeholders to quickly grasp key trends.

Univariate Analysis:

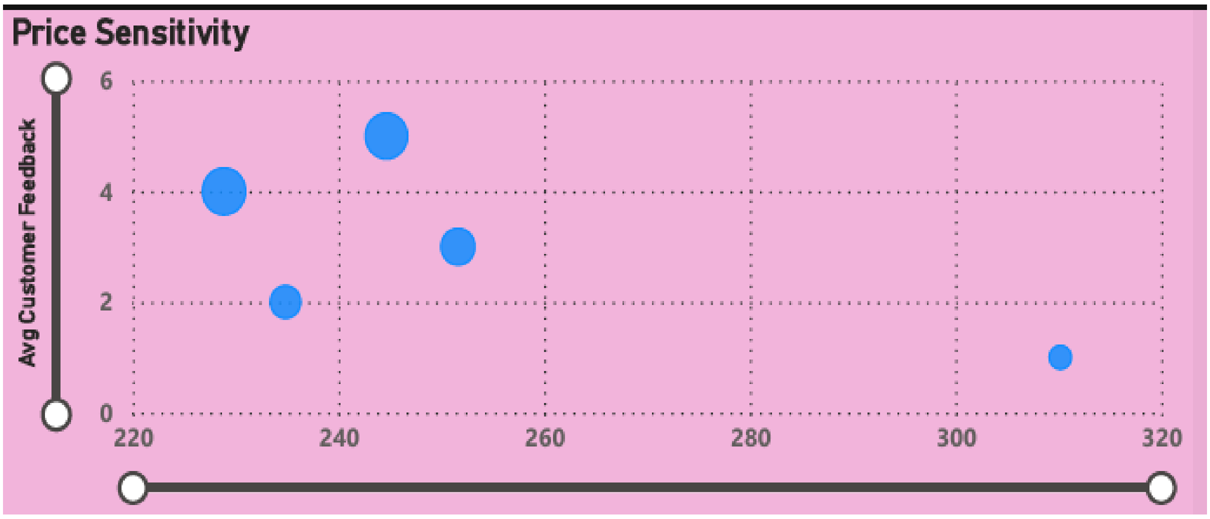
- Bar charts were used to represent total monthly sales across different disease categories, providing a clear comparison of sales performance.
- Pie charts were used to illustrate the distribution of sales based on insurance coverage, helping to identify trends in customer behavior.

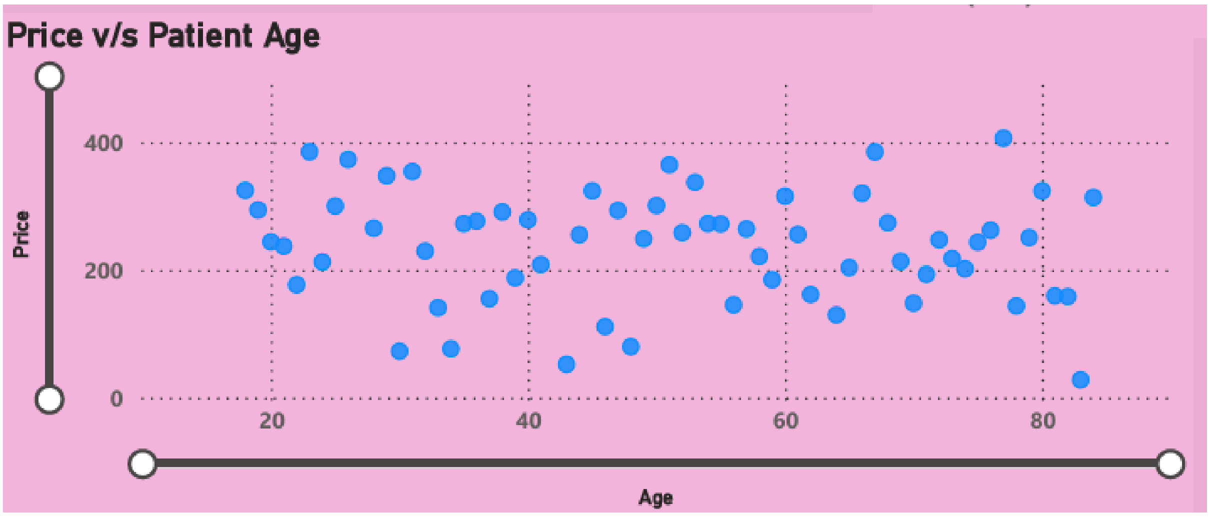




Bivariate Analysis:

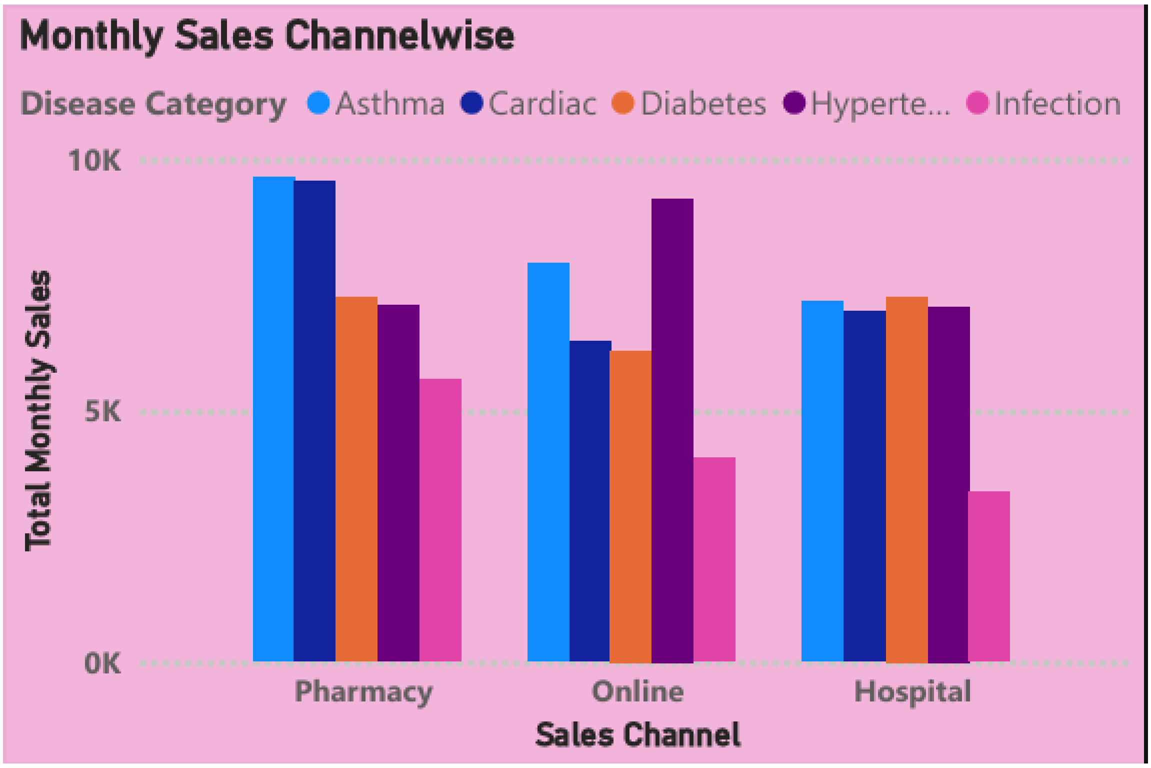
- Scatter plots were used to examine the relationship between price and customer feedback, highlighting the effect of pricing on customer satisfaction.
- Line charts were employed to explore how age correlates with purchasing decisions, revealing insights into the different needs of various age groups.





Multivariate Analysis:

- Stacked bar charts were used to compare sales across multiple sales channels for different disease categories, allowing for an in-depth view of how various factors interact.



These visualizations were selected to facilitate an understanding of the data from multiple perspectives, providing clear and actionable insights into customer behavior and sales performance.

8. Insights and Storytelling

Several key insights were drawn from the data:

- **Price Sensitivity:**
    - The highest customer satisfaction was observed in products priced between $240 and $260. This suggests that customers find this price range to be a sweet spot, balancing affordability and product value.
    - Products priced above $300 saw a significant drop in sales, indicating high price sensitivity among customers. This suggests that the company should consider revising the pricing strategy for higher-priced products.
- **Sales by Disease Category:**
    - Asthma and hypertension medications accounted for the highest sales, indicating a strong market demand in these categories.
    - On the other hand, infection-related products showed relatively low sales, possibly due to market saturation or lack of perceived value among customers.
- **Insurance Coverage Impact:**
    - A significant portion of the customers purchased products without insurance coverage, suggesting a reliance on out-of-pocket payments. However, customers with insurance coverage made more frequent purchases, highlighting the importance of affordability in driving sales.
- **Sales Channel Efficiency:**
    - Pharmacies were the dominant sales channel, followed by online platforms and hospitals.
    - Online platforms were especially effective for chronic disease categories, while hospitals were more successful for acute conditions, likely due to patient reliance on medical professionals for treatment decisions.

These insights were derived by examining the relationships between various data points and trends, providing a clear understanding of what drives customer behavior and sales.

---

9. Recommendations

Based on the insights generated, the following recommendations were made:

- **Pricing Strategy:**
    - Implement tiered pricing for products to cater to both budget-conscious customers and those willing to pay a premium for specialized treatments.
    - Explore subscription or value pack models for chronic disease treatments to encourage repeat purchases.
- **Marketing Focus:**

    - Increase promotional efforts for infection-related products, particularly in underperforming regions or markets.
    - Utilize digital platforms to target younger demographics who are more likely to engage in online purchasing.
- **Distribution Optimization:**
    - Expand the presence of online platforms to cater to the increasing trend of online shopping.
    - Strengthen pharmacy networks in underserved regions to ensure broader accessibility and improve market penetration.

These recommendations aim to optimize the pharmaceutical company's offerings, ensuring they are aligned with customer expectations while maintaining profitability.

---

10. Dataset Description

The dataset consists of 200+ rows and includes variables such as disease categories, price, customer age, insurance coverage, sales channels, and customer feedback. This variety of data points allows for a nuanced understanding of sales trends and customer behavior.

The structure of the dataset ensures that each variable is relevant to understanding different aspects of customer behavior, from the pricing sensitivity to the impact of insurance coverage and sales channels. By analyzing this dataset, pharmaceutical companies can refine their strategies to cater to diverse market segments.

Appendix

A. **Sample Data Table**

Field Name	Description
Prescription_ID	Unique identifier for each prescription.
Patient_ID	Unique identifier for patients (anonymized for privacy).
Age	Age of the patient.
Gender	Gender of the patient (e.g., Male, Female, Other).
Region	Geographic region of the patient (e.g., North, South, Urban, Rural).
Disease_Category	The medical condition being treated (e.g., Diabetes, Hypertension, Asthma).
Drug_Name	Name of the prescribed drug.
Drug_Category	Category of the drug (e.g., Antibiotic, Antidiabetic, Antiviral).
Dosage	Prescribed dosage (e.g., 500 mg, 10 mL).
Duration	Duration of the prescription in days.
Physician_ID	Unique identifier for the prescribing physician.

Physician_Specialization	Specialization of the prescribing physician (e.g., General Physician, Cardiologist).
Sales_Channel	Sales channel used (e.g., Pharmacy, Online, Hospital).
Price	Price of the drug.
Insurance_Coverage	Indicates if the drug was covered by insurance (Yes/No).
Purchase_Date	Date of purchase.
Customer_Feedback	Customer feedback on drug effectiveness (e.g., Scale of 1-5).
Side_Effects_Reported	Any side effects reported by the patient (e.g., Nausea, Headache).
Monthly_Sales	Monthly sales volume for the drug.
Competitor_Drug	Name of a competing drug, if any.