Tab 1

Final Project Report: New York Airbnb Open Data Analytics

Sector: Hospitality & Tourism Analytics

Group: 18 (Section B)

Team Members:

- **Pratiti** – Project Lead
- **Anuradha** – Data Lead
- **Jatin** – Analysis Lead
- **Shane** – Dashboard Lead
- **Priyansh**–   Strategy Lead
- **Chirag**– Quality Lead

Faculty/Institute: Newton School of Technology

1. Executive Summary

This project analyzes the New York Airbnb Open Data dataset to understand pricing behavior, host performance, room type distribution, and availability trends in the short-term rental market.

The dataset contained missing values, inconsistent formats, and pricing outliers. Through structured data cleaning and exploratory data analysis, the raw dataset was transformed into a reliable analytical model.

Key findings include:

- Manhattan listings command the highest prices.
- Entire homes generate higher revenue potential than private/shared rooms.
- A small percentage of hosts control a large share of listings.
- Price strongly varies by borough and room type.
- Availability patterns indicate demand concentration in central areas.

The project demonstrates how proper data cleaning and analysis can convert raw hospitality data into strategic business insights.

---

2. Sector & Business Context

Airbnb operates as a digital marketplace connecting property owners with travelers. In major cities like New York, short-term rentals compete with hotels and traditional accommodations.

Industry Characteristics:

- Dynamic pricing
- Strong location dependency
- Review-based reputation system
- High competition among hosts
- Seasonal and area-based demand variation

Current Challenges in Airbnb Data:

1.Missing price values
2.Inconsistent neighborhood naming
3.Duplicate host entries
4.Outliers in price (extremely high/low listings)
5.Null review counts
6.Solving these issues enables more accurate pricing strategies and host performance analysis

---

3. Problem Statement & Objectives

The Airbnb NYC dataset contains inconsistencies, missing values, and pricing outliers that make it difficult to extract reliable insights.

Formal Problem Definition

How can cleaned Airbnb listing data be used to understand pricing trends, host performance, and demand patterns across New York City?

Project Scope & Success Criteria

This project focuses on analyzing the **New York Airbnb Open Data** dataset to understand pricing patterns, location performance, host activity, and availability trends meeting the following criteria:

1 .Location Analysis – Borough and neighborhood performance
2..Pricing Analysis – Distribution, outliers, and affordability trends
3.Room Type Analysis – Performance comparison of property types
4..Host Analysis – Listing concentration and busiest hosts
5..Availability & Demand Analysis – Occupancy indicators

---

4. Data Description

The integrity of business intelligence rests upon the transparency of its data provenance and structural audit.

Technical Audit

- **Source:** https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata
- **Data Size:** Approximately 10,000 rows.
- **Structure:** 8 core columns containing temporal, categorical, and numeric dimensions.
- **Core Columns:**
    1. id
    2. Host_id
    3. neighbourhood_group
    4. neighbourhood
    5. latitude & longitude
    6. Room_type
    7. price
    8. minimum_nights
    9. number_of_reviews

Strategic Consideration: The dataset was intentionally provided in a "dirty" state, serving as a stress test for the data preparation workflows required before any "source of truth" can be established.

---

5. Data Cleaning & Preparation

All primary transformations were performed to improve data consistency, readability, and analytical reliability. The dataset was cleaned and standardized using spreadsheet-based preprocessing techniques.

Process Documentation

The following data preparation steps were carried out:

1. Duplicate Removal

Duplicate records were identified and removed to ensure each entry represented a unique observation. This prevented biased aggregations and incorrect statistical results during analysis.

2. Text Standardization

Textual fields were converted into Proper Case format to maintain consistency in naming conventions.
This helped avoid grouping errors caused by inconsistent capitalization (e.g., new york, New York, NEW YORK).

3. Irrelevant Column Removal

The columns House Rules and Last Review were removed as they were either unstructured textual data or not required for the analytical objectives.
This reduced dataset complexity and improved processing efficiency.

4. Column Renaming

Column names were modified to follow clear and readable naming conventions.
This improved dataset understandability and simplified formula writing and querying during analysis.

5. Boolean Conversion

Certain categorical columns were converted into Boolean values (True/False) to enable logical filtering and easier conditional analysis.

6. Ranking-Based Transformation

Selected variables were transformed into ranking-based values according to defined criteria.
This allowed better comparison, prioritization, and scoring during analytical evaluation.

---

6. KPI & Metric Framework

With data integrity secured, we turn to the KPI framework that aligns operational data with executive-level strategy.

KPI Name	Formula / Logic	Strategic Importance	Mapping to Objectives
Average Price (Borough-wise)	AVG(price) grouped by neighbourhood_group	Measures pricing benchmark across different boroughs and identifies premium markets.	Pricing Analysis
Total Listings	COUNT(id)	Measures pricing benchmark across different boroughs and identifies premium markets.	Market Structure
Listings by Room Type	COUNT(id) grouped by room_type	Identifies distribution of entire home,private rooms,and shared rooms listing.	Property Performance
Average Reviews per Listing	AVG(number_of_reviews)	Measure customer engagement and listing popularity.	Demand Analysis
Availability Rate	availability_365 / 365	Estimate occupancy level and booking frequency	Occupancy Insights

7. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand pricing patterns, occupancy behavior, and revenue drivers across NYC Airbnb listings.

The dataset was first cleaned by removing missing values, standardizing categorical fields, and converting price-related columns into numeric format. Additional calculated metrics were created, including **Occupancy Rate**, **Estimated Revenue**, and **Price Bands** to support pricing analysis.

Initial analysis showed:

- The **average market price** is approximately $619.
- The **average occupancy rate** is around 43%, indicating moderate demand consistency.
- Revenue varies significantly across **room types and neighbourhood groups**, with Shared Rooms and Private Rooms showing strong revenue efficiency.
- Manhattan has higher pricing levels, while Brooklyn demonstrates a strong balance between occupancy and revenue.
- Correlation analysis between **price and occupancy** showed a very weak negative relationship, suggesting that demand is influenced by multiple factors beyond pricing alone.

Overall, the EDA revealed that revenue optimization depends on a combination of **location, room type, and pricing strategy**, forming the foundation for the pricing recommendations developed in this study.

---

8. Advanced Analysis

Advanced segmentation moves beyond "what happened" to provide root-cause understanding of market structure and demand behavior within the Airbnb ecosystem in New York City.

- **Customer Segmentation (Entire Home vs. Private Room):**Data reveals that guests booking “Entire home/apartment” listings exhibit higher average spending compared to those booking private or shared rooms. We hypothesize this is driven by traveler preference for privacy, group accommodation needs, and higher perceived value, which increases willingness to pay.

- **Location Premium Effect (Manhattan vs. Outer Boroughs):**We investigated the significant price gap between Manhattan and other boroughs. Root-cause analysis indicates this is driven primarily by proximity to commercial centers, tourist attractions, and transportation hubs, rather than differences in listing quality alone. Location acts as the dominant pricing lever.
- **Host Concentration Impact:**Analysis shows that a small percentage of hosts manage multiple listings. Root-cause interpretation suggests partial professionalization of the platform, where property managers operate multiple units, leading to higher operational efficiency and stronger market presence.

9. Dashboard Design

The Google Sheets dashboard democratizes data access, allowing non-technical stakeholders to interact with high-value insights derived from the Airbnb New York dataset.

Architecture & Objective

The dashboard provides a consolidated view of pricing trends, listing distribution, host performance, and demand indicators across New York City.

- **View Structure:** A centralized "Executive Overview" presents key KPIs such as Average Price, Total Listings, Median Price, and Availability Rate. A secondary "Operational Drill-down" section enables borough-wise, room-type-wise, and host-level performance analysis.
- **Interactive Filters:**Integrated slicers for Borough (neighbourhood_group), Room Type, Price Range, and Minimum Nights allow real-time scenario testing and dynamic comparison across market segments.

- **Implementation Logic:** Dynamic Pivot Tables powered by **SUMIFS, COUNTIFS, AVERAGEIFS,** and conditional logic formulas ensure that the dashboard automatically updates when new listing data is added. Structured data formatting guarantees consistency and scalability, making the dashboard a "living" analytical tool rather than a static report.

---

10. Insights Summary

The following 10 insights are presented in "decision language" to facilitate immediate strategic intervention within the Airbnb market in New York City:

1. **Location Revenue Concentration:**
   A small cluster of Manhattan neighborhoods drives a disproportionate share of total listing value, while outer borough areas contribute significantly lower average pricing.
2. **Entire Home Premium:**
   Entire home/apartment listings generate substantially higher average prices compared to private and shared rooms, positioning them as primary revenue drivers.
3. **Host Market Concentration:**
   A limited percentage of hosts manage multiple listings, indicating partial professionalization and competitive dominance by multi-property operators.
4. **Price Skewness Impact:**
   Luxury outlier listings inflate the overall average price; median pricing provides a more accurate market benchmark.
5. **Demand Density Effect:**
   Listings with lower availability_365 values consistently show higher review counts, suggesting stronger occupancy in premium locations.
6. **Brooklyn Growth Signal:**
   Brooklyn demonstrates high listing volume with competitive pricing, indicating strong mid-tier demand positioning.
7. **Minimum Nights Barrier:**
   Listings with higher minimum night requirements experience reduced review frequency, suggesting booking friction for short-term travelers.

8. **Data Cleaning Variance:**
   Pre-cleaned data contained duplicate entries and formatting inconsistencies that could distort pricing analysis if left uncorrected.
9. **Budget Segment Stability:**
   Private rooms maintain stable demand among price-sensitive travelers, especially in non-central boroughs.
10. **Location as Primary Price Lever:**
    Geographic location exerts stronger influence on pricing than room type alone, confirming neighborhood-level demand as the dominant pricing determinant.

---

11. Recommendations

Recommendation	Linked Insight	Projected Business Impact
Dynamic Pricing by Borough	Location Revenue Concentration	**High:** 10–15% potential revenue increase by aligning prices with neighborhood demand intensity.
Entire Home Revenue Optimization	Entire Home Premium	**High:** 8–12% improvement in listing revenue by prioritizing high-performing property types.
Flexible Minimum Night Policy	Minimum Nights Barrier	**Medium:** 6–10% increase in booking frequency by reducing stay restrictions.
Host Performance Program	Host Market Concentration	**Medium:** Improved listing quality and competitive balance through performance benchmarking and incentives.

---

12. Impact Estimation

The following estimates justify the analytical investment:

- **Revenue Optimization:**Target a 10–15% increase in listing revenue through dynamic borough-based pricing and room-type optimization strategies.
- **Improved Occupancy Efficiency:**Projected 8–12% improvement in booking rates by reducing restrictive minimum night policies and optimizing availability schedules.
- **Market Positioning Enhancement:**Strategic focus on high-demand neighborhoods such as Manhattan and Brooklyn can strengthen competitive positioning and increase visibility in premium segments.
- **Operational Clarity:**Standardized data cleaning and structured KPI reporting eliminate pricing distortions caused by outliers and duplicates, reducing analytical error and preventing flawed strategic decisions.

---

13. Limitations:

- **Data Quality Constraints:**
  The original dataset contained missing values, formatting inconsistencies, and pricing outliers. Cleaning required statistical assumptions (e.g., median imputation, outlier trimming), which may slightly affect precision-level accuracy.
- **Temporal Limits:**
  The dataset represents a static snapshot of listings and does not include multi-year historical trends. As a result, long-term forecasting and seasonal pattern validation remain limited.
- **External Market Factors:**
  The dataset excludes key external variables such as tourism seasons, competitor pricing, local regulations, economic conditions, and special events, all of which significantly influence Airbnb demand and pricing.
- **Revenue Proxy Limitation:**
  Actual booking revenue data was not available. Price and availability were used as proxy indicators for demand and performance.

---

14. Future Scope

This project serves as the foundation for a more advanced intelligence roadmap:

- **Predictive Pricing Modeling:**
  Moving from descriptive analysis to machine-learning-based price prediction using location, reviews, availability, and room type as features.
- **Demand Forecasting:**
  Incorporating seasonal trends and external tourism data to predict occupancy cycles and optimize pricing strategies.
- **Sentiment Analysis on Reviews:**
  Applying Natural Language Processing (NLP) techniques to analyze guest reviews for quality insights and service improvement.
- **Host Performance Scoring Model:**
  Developing a composite performance index combining reviews, pricing efficiency, and occupancy metrics.
- **Automated Data Pipeline:**
  Integrating Airbnb API or automated data ingestion workflows to eliminate manual CSV updates and maintain real-time dashboard accuracy.

---

15. Conclusion

This project successfully demonstrates the transformation of raw Airbnb listing data into a high-value strategic asset. By applying systematic data cleaning, structured KPI development, and analytical segmentation, we developed a clear understanding of pricing behavior, host concentration, and demand dynamics within New York City.

The transition from unstructured listing data to actionable business intelligence highlights the importance of data-driven decision-making in modern hospitality markets. With further predictive enhancements and automation, this analytical framework can evolve into a scalable pricing and demand optimization system for Airbnb hosts and market analysts.

16. Appendix

Data Dictionary

1. **id:** Unique alphanumeric identifier for each Airbnb listing.
2. **host_id:** Unique identifier assigned to each host.
3. **host_name:** Name of the host managing the listing (Categorical).
4. **neighbourhood_group:** Borough in which the listing is located (Manhattan, Brooklyn, Queens, Bronx, Staten Island).
5. **neighbourhood:** Specific neighborhood within the borough (Categorical).
6. **latitude:** Geographical latitude coordinate of the listing (Numeric).
7. **longitude:** Geographical longitude coordinate of the listing (Numeric).
8. **room_type:** Type of accommodation offered (Entire home/apartment, Private room, Shared room).
9. **price:** Nightly rental price of the listing (Currency).
10. **minimum_nights:** Minimum number of nights required per booking (Numeric).
11. **number_of_reviews:** Total number of reviews received by the listing (Numeric).
12. **last_review:** Date of the most recent review (Date).
13. **reviews_per_month:** Average number of reviews received per month (Numeric).
14. **availability_365:** Number of days the listing is available for booking in a year (Numeric).

17. Contribution Matrix

Transparency and collaborative integrity are fundamental to this project's success.

Team Member	Dataset & Sourcing	Cleaning	KPI & Analysis	Dashboard	Report Writing	PPT	Overall Role
Pratiti	✔		✔	✔			**Project Lead**
Anuradha	✔	✔					**Data Lead**
Jatin	✔				✔		**Analysis Lead**
Shane	✔			✔			**Dashboard Lead**
Priyansh	✔			-		✔	**Strategy Lead**
Chirag	✔					✔	**Quality Lead**

Declaration: We confirm that the above contribution details are accurate and verifiable through Google Sheets version history and submitted artifacts.

Tab 2

FINAL PROJECT REPORT

Cafe Analytics Dashboard

Enterprise-Grade Retail Intelligence | Data Visualization & Analytics Capstone Project

1. Executive Summary

In the modern Food & Beverage (F&B) industry, operational success is directly linked to the ability to convert high-volume transactional data into actionable business intelligence. This project presents the transformation of approximately **10,000 raw Point-of-Sale (POS) records into 6,658 clean, validated transactions**, enabling accurate analysis and strategic decision-making.

The raw dataset suffered from critical data quality issues including:

- Missing and inconsistent values
- Incorrect revenue calculations
- Duplicate records
- Non-standard formatting

A structured data cleaning pipeline (implemented in Google Sheets and aligned with documented methodology) ensured:

- **100% financial accuracy**
- **0 missing values**
- **0 duplicates**
- Fully standardized dataset

Key Insights

- **In-store transactions contribute ~70% of total revenue**, indicating heavy dependence on physical operations
- **Cash remains the dominant payment method (~39%)**, though digital and card payments collectively represent a strong alternative
- **Revenue follows a stable monthly trend**, indicating predictable demand patterns
- **Product performance follows a Pareto distribution**, where a limited number of items drive the majority of revenue

Strategic Recommendations

- Optimize menu using **high-performing product focus**
- Promote **digital payments to increase efficiency**
- Improve **takeaway channel performance**
- Introduce **bundling strategies to increase average order value**

🎯 Business Outcome

This project enables transition from:

> ❌ Data storage → ✅ Data-driven decision-making

---

2. 🏢 Sector & Business Context

The F&B sector operates under:

- High transaction frequency
- Low profit margins
- Dynamic demand patterns

Each transaction captures multiple dimensions:

- Item purchased
- Quantity
- Payment method
- Service mode
- Time

🚨 Core Challenges

Challenge	Business Impact
Inaccurate revenue data	Misleading financial decisions
Missing categorical values	Broken segmentation
Duplicate transactions	Inflated KPIs
Inconsistent formats	Analysis inefficiency

💡 Why This Problem Matters

Without clean data:

- No reliable KPIs
- No forecasting
- No optimization

With clean data:

- Clear insights
- Better decisions
- Scalable operations

---

3. 🎯 Problem Statement & Objectives

Problem Statement

> “How can raw POS data be transformed into accurate, structured insights to identify revenue drivers, customer behavior, and operational inefficiencies in a cafe business?”

Objectives

- Build a **clean, validated dataset**
- Develop a **KPI-driven analytics framework**
- Design an **interactive dashboard**
- Generate **actionable business insights**

Success Criteria

- 100% data accuracy
- Zero missing/duplicate records
- Business-ready insights

---

4. 📂 Data Description

Source

Kaggle – *Cafe Sales Dirty Dataset*

Dataset Overview

Metric	Value
Raw Records	~10,000
Cleaned Records	6,658
Columns	9
Time Period	Jan–Dec 2023

Schema

Column	Description
transaction_id	Unique transaction identifier
item	Product purchased
quantity	Units sold
unit_price	Price per unit
total_spent	Revenue
payment_method	Cash / Card / Wallet
location	In-store / Takeaway
transaction_date	Date
transaction_month	Month

Limitations

- No customer-level data
- No cost/margin data
- Single-year dataset

5. 🧹 Data Cleaning & Preparation

(Aligned with full cleaning documentation)

🔧 Key Cleaning Steps

- Recalculated revenue → **Total = Quantity × Price**
- Corrected invalid price entries
- Removed rows with invalid quantities
- Eliminated **duplicate records (118 rows)**
- Standardized categorical values
- Removed whitespace inconsistencies
- Handled missing values
- Added derived feature → **Transaction Month**

📊 Impact

Metric	Before	After
Records	~10,000	6,658
Null Values	1000+	0
Revenue Errors	High	0%
Duplicates	118	0

✅ Outcome

Dataset is **fully validated, audit-ready, and analytics-grade**

---

6. 📊 KPI & Metric Framework

KPI	Formula	Business Value
Total Revenue	SUM(total_spent)	Measures overall performance
Avg Order Value	Revenue / Orders	Measures spending behavior

Orders	Count of transactions	Demand indicator
Revenue by Item	SUM per item	Product performance
Payment Share	% revenue by method	Customer preference
Revenue by Location	SUM by location	Channel performance

🎯 Why KPIs Matter

They directly answer:

- What drives revenue?
- How customers behave?
- Where optimization is needed?

7. 📈 Exploratory Data Analysis (EDA)

🔍 Monthly Trend

Revenue demonstrates **low volatility**, indicating:

- Stable demand
- Predictable operations

🔍 Product Demand

- Demand concentrated in a few products
- Indicates **Pareto distribution (80/20 behavior)**

🔍 Service Mode

- In-store dominates revenue
- Takeaway significantly underutilized

🔍 Payment Behavior

- Cash leads individually
- Digital + card combined are comparable → shift opportunity

---

8. 🧠 Advanced Analysis

Root Cause Insights

Observation	Root Cause
Low takeaway revenue	Limited promotion & UX
Payment imbalance	No digital incentives
Revenue concentration	Limited menu optimization

Key Interpretation

- Business is **stable but not optimized**
- Revenue heavily dependent on **few drivers**
- Untapped growth exists in **digital + takeaway channels**

---

9. 📊 Dashboard Design

🎯 Objective

Provide **real-time decision support for managers**

Dashboard Structure

Executive View

- Orders → 20,107
- Revenue → $60,484
- AOV → $9

Operational View

- Monthly trends
- Product demand
- Payment distribution
- Service mode comparison

⚙️ Features

- Interactive filters (Item, Month, Location, Payment)
- Pivot-based aggregation
- Dynamic charts

💡 Decision Enablement

Managers can:

- Identify top products
- Track performance trends
- Compare payment behavior
- Optimize operations instantly

---

10. 🔥 Key Insights (Decision Language)

1. Revenue is heavily concentrated in a limited number of products, indicating dependency risk
2. In-store contributes majority revenue, limiting scalability
3. Takeaway channel is underperforming relative to its potential
4. Payment distribution is balanced, but digital adoption is not maximized
5. Monthly sales stability indicates predictable demand cycles
6. Product demand consistency supports inventory planning
7. Revenue structure suggests opportunity for bundling strategies
8. Lack of volatility indicates operational stability but limited growth initiatives

---

11. 🚀 Recommendations

Recommendation	Insight	Impact

Promote top products	Revenue concentration	↑ Revenue
Introduce combo offers	AOV opportunity	↑ Basket size
Incentivize digital payments	Payment imbalance	↑ Efficiency
Improve takeaway UX	Underperformance	↑ Channel revenue
Optimize menu	Low performers	↓ Waste

12. 📈 Impact Estimation

Area	Impact
Revenue	+10–15%
Cost	-10%
Efficiency	+15%
Risk	Reduced

13. ⚠️ Limitations

- No customer segmentation
- No profitability analysis
- Limited time scope

14. 🔮 Future Scope

- ML-based demand forecasting
- Customer segmentation
- Real-time dashboard integration
- POS automation

15. ✅ Conclusion

This project successfully demonstrates:

> **Raw Data → Clean Data → Insights → Strategic Decisions**

The transformation from unstructured POS data to an **enterprise-grade analytics dashboard** provides a scalable framework for data-driven decision-making in the F&B sector.

---

16. 📎 Appendix

Includes:

- Data dictionary
- Cleaning logic
- Pivot calculations
