Introduction  
This report reviews a exploratory and predictive analysis of an NCR ride-hailing dataset (150,000 rows, 21–24 columns) focused on booking outcomes, cancellations, timings, vehicle types, locations, payment methods, ratings, and a baseline ML classification of booking status. The notebook performs data loading, cleaning with extensive NA imputation, feature engineering, EDA visualizations, and three classifiers (Logistic Regression, KNN, Decision Tree) with reported accuracies, enabling operational insights on demand patterns, cancellations, and payment-method performance.

Data loading and structure

* The analysis loads ncr\_ride\_bookings.csv into a Data Frame with 150,000 entries and 21 columns initially, expanding to 24 after engineering; key fields include Date, Time, Booking Status, Vehicle Type, locations, Avg VTAT/CTAT, cancellation flags and reasons, booking value, distance, ratings, and payment method.
* Non-null overview shows heavy sparsity in cancellation details and outcomes: Avg CTAT has 48,000 missing, Booking Value/Distance 48,000 missing each, ratings 57,000 missing each, and cancellation-reason columns mostly missing; after imputation and type conversions, Date and Time become datetime64, numeric fields are fully populated, and new features is\_cancelled, DayOfWeek, and Hour are added to reach 24 columns.

Data quality and imputation

* Missing-value counts reveal that cancellation indicators and reasons are extremely sparse relative to the population: Cancelled Rides by Customer 139,500 missing, Cancelled Rides by Driver 123,000 missing, Incomplete Rides 141,000 missing; ratings and monetary metrics also have large gaps.
* Imputation strategy: fill NA for categorical reasons with “Not cancelled”/“Not incomplete”, cancellation flags with 0, financial/travel metrics with 0, ratings with 0, and Avg VTAT/CTAT with 0; payment method NA filled with “Cancelled or incomplete ride”. This enables complete-case modeling but risks label leakage and distributional distortion where zero is non-physical (e.g., 0 km distance).

Feature engineering and target definition

* DayOfWeek and Hour are derived from Date and Time, enabling time-based grouping for EDA on demand and cancellations, and a combined is\_cancelled indicator is defined as 1 if either customer or driver cancellation equals 1, otherwise 0.
* Class balance for is\_cancelled shows 112,500 not-cancelled vs 37,500 cancelled, implying a 25% cancellation rate; acceptance rate is reported as 75.00% and “success rate” as 94.00% based on incomplete=0 for rides completed, suggesting incompletions are 6% of all observations after imputations.

Descriptive statistics

* Booking Value averages 508.3 with high variance (std ~395.8) and ranges from 50 to 4,277, aligning with a long-tailed fare distribution typical of variable trip lengths and vehicle types.
* Ride Distance averages 24.64 km (std ~14.00) spanning 1 to 50 km, and ratings cluster positively: driver mean 4.23 and customer mean 4.40 on 3–5 scales, with medians at 4.3/4.5 respectively.

Payment methods and counts

* Payment methods in the data show UPI as the largest share (45,909), followed by Cash (25,367), Uber Wallet (12,276), Credit Card (10,209), and Debit Card (8,239), indicating digital-first usage with a substantial cash segment.
* EDA includes a stacked proportion chart cross-tabulating Payment Method by Booking Status and a count plot by status, suitable to evaluate if cancellation or completion rates vary by tender type; these visualizations support payment-policy and incentive adjustments.

Booking status overview and time trends

* A count plot summarizes frequency of each Booking Status, with labels including Completed, Incomplete, and No Driver Found; this distinguishes supply constraints vs rider-driven failure states.
* A line plot of daily counts by status shows temporal variation across Date for each status, enabling detection of spikes in “No Driver Found” or “Incomplete” which can relate to events, weather, or supply shifts.

Cancellations analysis

* Overall cancellation mechanics are split into customer vs driver cancellations; pies quantify proportions, and bar charts show cancellations by day of week and by hour derived from is\_cancelled\_customer to highlight peak-risk periods.
* A bar chart of “Reasons for cancelling by Customer” excludes “Not cancelled” and ranks actual reasons; this provides direct levers for product fixes and communications, though caution is needed given sparse reason labeling.

Vehicle types and booking outcomes

* Vehicle Type distribution is visualized with a count plot, confirming mix across Bike, Auto, Go Mini/Go Sedan, Premier Sedan, eBike, etc., facilitating supply planning and segmentation.
* Crosstab of Vehicle Type vs Booking Status exposes which categories over-index in cancellations or “No Driver Found”; an additional bar of customer cancellation rate by vehicle type (using mean of boolean) reveals per-type risk differentials.

Locations

* Top 10 pickup locations by count highlight hotspots such as MG Road, Golf Course Road, and key NCR neighborhoods; these areas merit dedicated driver supply, dynamic pricing, and geofencing considerations to sustain completion rates.
* Using these rank plots, operations can tune supply alignment to demand, especially during identified high-cancellation hours and days to reduce “No Driver Found” and negative customer experiences.

Time-of-day economics

* Average Booking Value by Hour is computed and plotted as a line, indicating intra-day fare variation that likely correlates with trip length and congestion patterns; this informs time-based promotions or supply incentives.
* Cancellations by Hour further show when riders abandon trips or drivers cancel, enabling targeted nudges, ETA estimates tuning, or surge logic adjustments during vulnerable windows.

Ratings relationships

* A heatmap of Driver Ratings x Customer Rating cross-tab reveals co-distribution patterns; higher driver ratings tend to co-occur with higher customer ratings, consistent with reciprocal satisfaction dynamics.
* A top-5 correlation heatmap (absolute-sum heuristic) surfaces numeric features with stronger interrelations; while not causal, it directs feature selection and regularization choices for modeling.

Target and label considerations

* The analysis treats Booking Status as the classification target while also engineering is\_cancelled from cancellation flags; this duality suggests alternative targets like binary completion/cancellation for cleaner modeling.
* Because multiple downstream plots and the ML step encode categorical columns with label encoding, the ordinal imposition on nominal features could mislead distance-based models and linear terms; one-hot encoding would be preferable for non-ordered categories.

Machine learning setup

* Features X drop Date, Time, Hour, Booking Status, and Booking ID while retaining engineered and categorical variables; y is the label-encoded Booking Status, and numeric columns are imputed with median while categoricals are label-encoded as strings.
* A stratified 80/20 split with random\_state=42 is used; three models are trained: Logistic Regression, KNN, and a shallow Decision Tree with max\_depth=2 to reduce variance and improve interpretability.

Model performance

* Reported test accuracies: Logistic Regression 0.7956, KNN 0.7570, and Decision Tree 0.8700, indicating the shallow tree outperforms baselines on this encoding and feature space.
* No confusion matrices or class-by-class metrics are printed; given possible class imbalance across multiple Booking Status labels, adding precision/recall/F1 and macro/micro averaging would clarify real-world performance beyond accuracy.

Key insights and operational recommendations

* Cancellation burden: A 25% cancellation rate via is\_cancelled underscores material friction; driver supply balancing and pre-acceptance ETA/price transparency could reduce rider cancellations, while driver-side incentives and cancellation penalty policies could reduce driver cancellations.
* Incompletion vs acceptance: With a 94% “success” on incomplete=0 but only 75% “acceptance,” many failures occur pre-ride; focus upstream on matching speed (VTAT/CTAT), demand-supply alignment by hour/day/location, and early-channel messaging to deflect churn before dispatch.
* Vehicle-type risk: Vehicle categories show different cancellation propensities; prioritize reliability in higher-risk types via targeted driver incentives during peak hours or dynamic reassignment to more available classes.
* Payment behavior: UPI dominance suggests digital readiness; review stacked proportions to detect if cash or wallet transactions correlate with higher non-completion, then adjust payment option defaults and promotions to favor lower-risk methods.
* Time targeting: Cancellations concentrate at specific hours and days; schedule driver incentives and surge multipliers for those windows in top pickup locations to reduce “No Driver Found.”
* Data remediation: Replace zero-filled metrics (distance, value, ratings) for missing post-cancellation entries with NA and exclude from modeling of completed-ride targets; adopt explicit masks to avoid bias from structural zeros.

Conclusion  
The analysis provides a solid EDA foundation, revealing a 25% cancellation rate, strong digital payment usage, distinct vehicle-type and time-of-day patterns, and promising baseline classification results led by a shallow decision tree at 87% accuracy. To progress from descriptive to actionable, refine imputation, adopt robust encoding and evaluation, clarify the target (pre-ride cancellations vs ride completion), and integrate explainability—steps that will yield reliable levers for reducing cancellations, improving matching, and optimizing supply across NCR hotspots and peak periods.