Data Analysis Final Assignment Report

Team: Current Catalysts
Vikram Anand & $UmerFarooq$

29.01.2026

1 Contributions

- Umer Farooq:
  - Dataset selection and acquisition
  - Data quality analysis and preprocessing
  - Probability analysis taskspipeline
- Vikram Anand:
  - Regression modeling and interpretation
  - Report writing and figure polishing
  - Visualizations and EDA

2 Dataset Description

- Dataset name and source: Bike Sharing
- Why it is suitable for time-series analysis: The Bike Sharing dataset shows bike rental counts over every hour for a period of 2 years. Changes in bike rentals help visualize civic activity in a major city (Washington DC).
- Time period covered and sampling frequency: data points available from January 2011-December 2012. Sampled every hour.
- Key variables analyzed (signals, sensors, physical quantities): Number of rentals, Temperature, Day of Week, Season, humidity, and wind speed.
- Size and structure:
  - Number of observations (rows):17,379
  - Number of features (columns):18
  - Target variable(s) if any: cnt (Bike rental count at a particular hour)
- Missing data summary: Missing timestamps were detected by building a continuous hourly index and counting absent hours (165 missing timestamps)
- Any known limitations or caveats: Observations are temporally dependent (hourly autocorrelation), which slows convergence in sampling-based results and violates IID assumptions for some statistical tests.

3 Task 1. Data Preprocessing and Basic Analysis

3.1 Basic statistical analysis using pandas

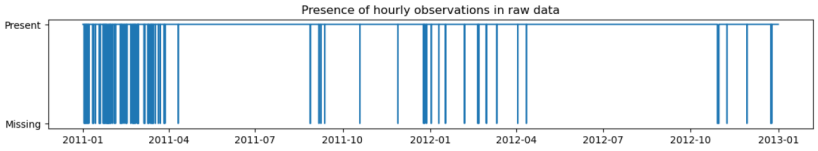
- Descriptive stats (mean, std, min, max, quantiles) for key variables:
  - The statistical summary of key features is as follows:

	count	mean	std	min	25%	50%	75%	max
cnt	17379.0	189.463088	181.387599	1.00	40.0000	142.0000	281.0000	977.0000
temp	17379.0	0.496987	0.192556	0.02	0.3400	0.5000	0.6600	1.0000
atemp	17379.0	0.475775	0.171850	0.00	0.3333	0.4848	0.6212	1.0000
hum	17379.0	0.627229	0.192930	0.00	0.4800	0.6300	0.7800	1.0000
windspeed	17379.0	0.190098	0.122340	0.00	0.1045	0.1940	0.2537	0.8507
casual	17379.0	35.676218	49.305030	0.00	4.0000	17.0000	48.0000	367.0000
registered	17379.0	153.786869	151.357286	0.00	34.0000	115.0000	220.0000	886.0000

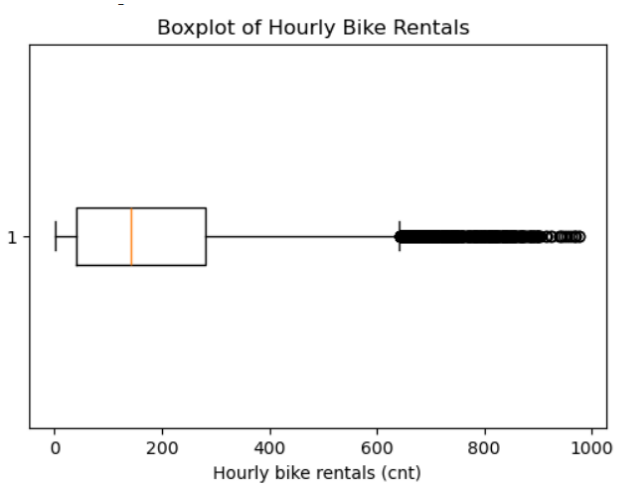
- Grouped summaries where relevant (by day, device, category, test run):
  Bike rental count by hour of day: showed peak hours of renting between hours 10-16.
  By day: Weekdays had more overall rentals than weekends.
  By Season: fall season had the highest among all the 4 seasons.

3.2 Original data quality analysis including visualization

- Missingness patterns (counts, heatmap, timeline gaps): 165 Timeline gaps were observed and represented by a timeseries plot.



- Outliers and suspicious values (plots and rule used):



  Although infrequent rental counts were observed, they were used as valid data to model factors influencing bike rentals.

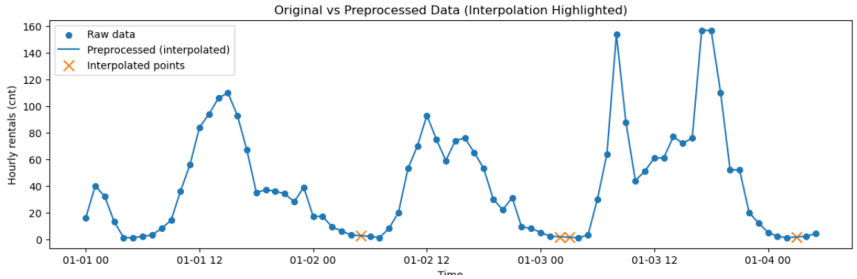
- Consistency checks (timestamps order, duplicates, impossible values):No duplicates observed.

3.3 Data preprocessing

- Cleaning steps performed:Time method of interpolation was used.
- Missing-value treatment (drop, impute, interpolate, forward fill, etc.): Time method was used to interpolate missing rows.
- Outlier handling (range, threshold, IQR, percentile, justify choice):Outliers were not considered for this dataset.
- Feature engineering (e.g., scaling/normalization, log):Features such as temperature, humidity and windspeed were already normalized.
- Final dataset shape after preprocessing:12 columns & 17354 rows

3.4 Preprocessed vs original data visual analysis

- Before vs after comparison plots (at least 2 to 3 key variables): The dataset was already preprocessed. Thus rows with missing timestamps were added and interpolated.

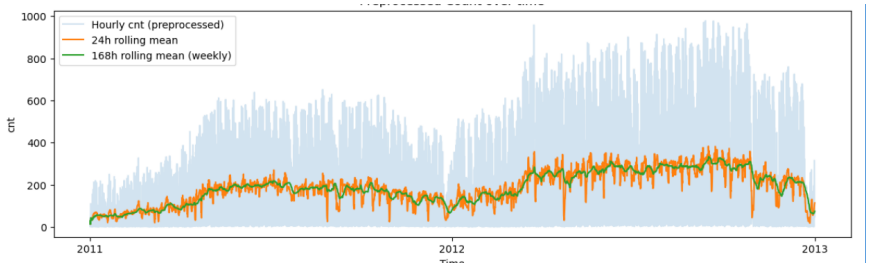


- What improved and what trade-offs exist: No unrecorded periods in the dataset at the cost of artificially smoothened time-series from time interpolation

4 Task 2. Visualization and Exploratory Analysis

4.1 Time series visualizations

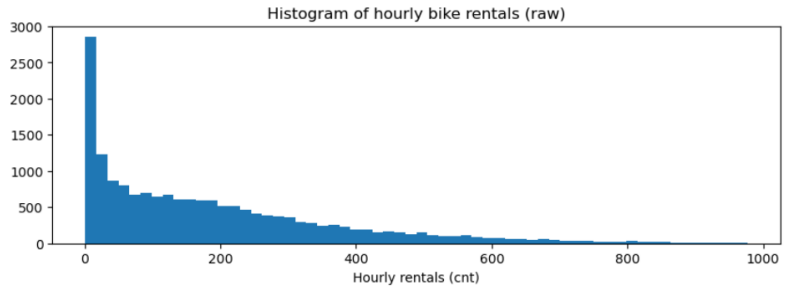
- Plot of main variable(s) over time:



- Annotations for notable events or pattern shifts (if applicable):

4.2 Distribution analysis with histograms

- Histograms for key numeric variables:

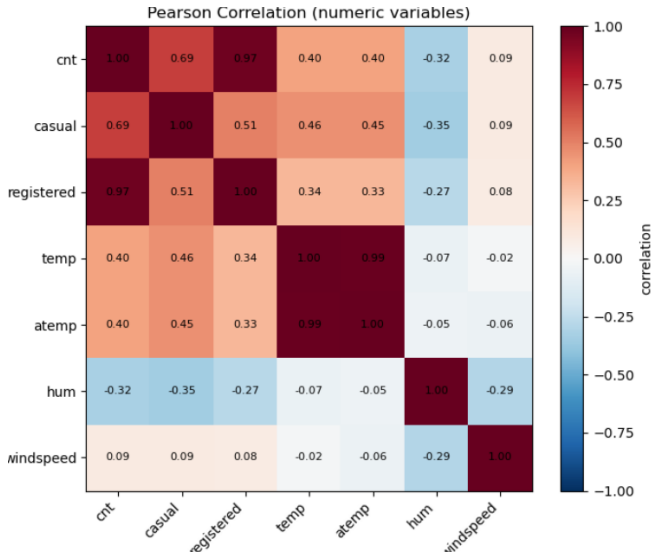


- Notes on skewness, heavy tails, multi-modality: Data is right skewed as most hours have low rentals. Heavy tails observed during certain hours.

4.3 Correlation analysis and heatmaps

- Correlation type used (Pearson or Spearman) and why:Pearson correlation was used to measure linear relationship between variables.This was used because the dataset focuses on the linear relationship between bike demand and other variables.

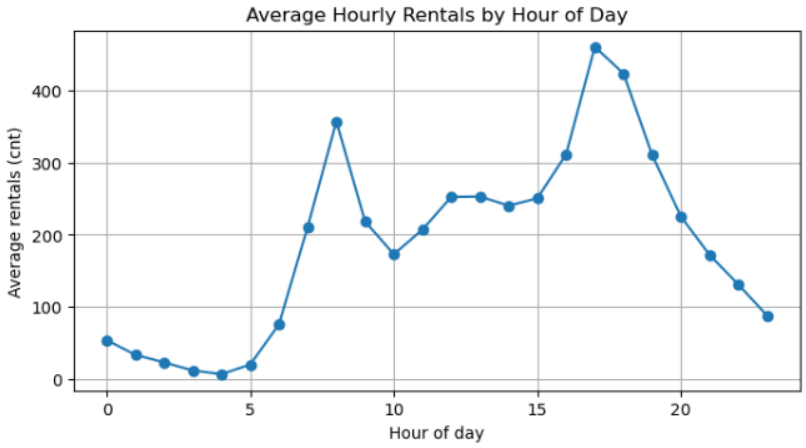
- Heatmap and top correlated pairs with short interpretation: Bike rentals show moderate correlation with temperature(0.40).



4.4 Daily pattern analysis

- Aggregation method (hourly means, day-of-week, rolling averages): Rental data was aggregated through rolling mean of rentals throughout a day and a week.

- Plots showing daily cycles or weekday-weekend differences:



- What patterns are stable vs noisy: The 24 hour rolling averages are more noisy and the weekly rolling average is more stable.

4.5 Summary of observed patterns, similar to True/False questions

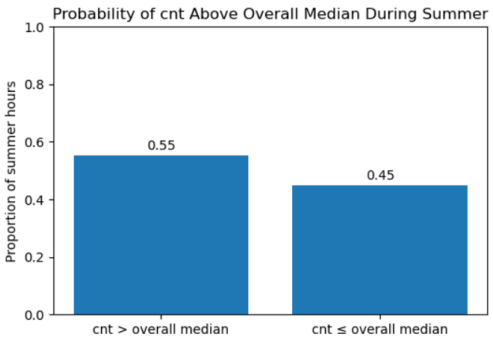
Write short, testable statements and answer them based on evidence. Example format below. The mean bike rentals are 189.4

- Peak Hours between 7-9; Evidence:Mean(263)
- Peak Hours between 16-18; Evidence:Mean(398)
- Weekdays are more active than Weekends; Evidence:Mean(189)

5 Task 3. Probability Analysis

5.1 Threshold-based probability estimation

- Define threshold(s) and justify choice: The threshold is the median of hourly rentals because it defines the middle point of rental counts.
- Estimate probabilities of exceeding thresholds:55% of summer hours had rentals above the threshold.
- Visual support (e.g., empirical CDF, bar plot, timeline highlights):



5.2 Cross tabulation analysis

- Define two categorical variables (or binned numeric variables):Weekday vs Count larger than median
- Present contingency table and interpret key cells:

```
Cross-tab: Weekday × (cnt > median)
cnt_above_median  False  True
is_weekday
False             0.538  0.462
True              0.486  0.514
```

46% of Weekends had rentals larger than the median count.

5.3 Conditional probability analysis

- Define events A and B: A: Rentals count above median. B: Current season is Summer
- Compute and interpret $P(A)$, $P(B)$, $P(A \mid B)$, $P(B \mid A)$: P(A)=0.499, P(B)=0.251, $P(A \mid B)$=0.552, $P(B \mid A)$=0.278
- Include at least one meaningful comparison and conclusion: Rentals above median are more likely to occur on a summer day than on a randomly selected day.

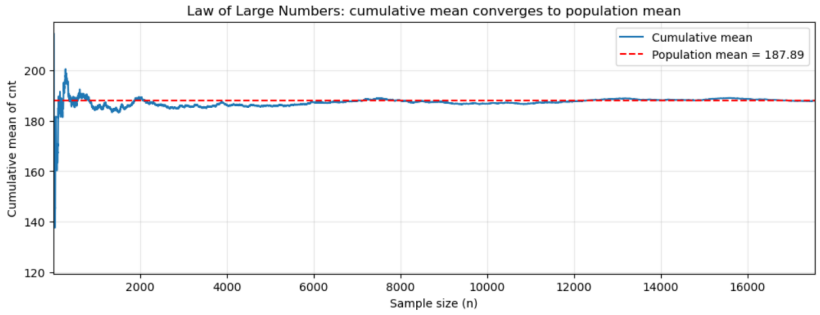
5.4 Summary of observations from each probability task

- Key takeaway from threshold probability: Summers hours get have typically more rentals.
- Key takeaway from crosstab: Weekends see lesser peak demand hours
- Key takeaway from conditional probability: Summer hours exceed typical demand but high demand is not exclusive to Summer.

6 Task 4. Statistical Theory Applications

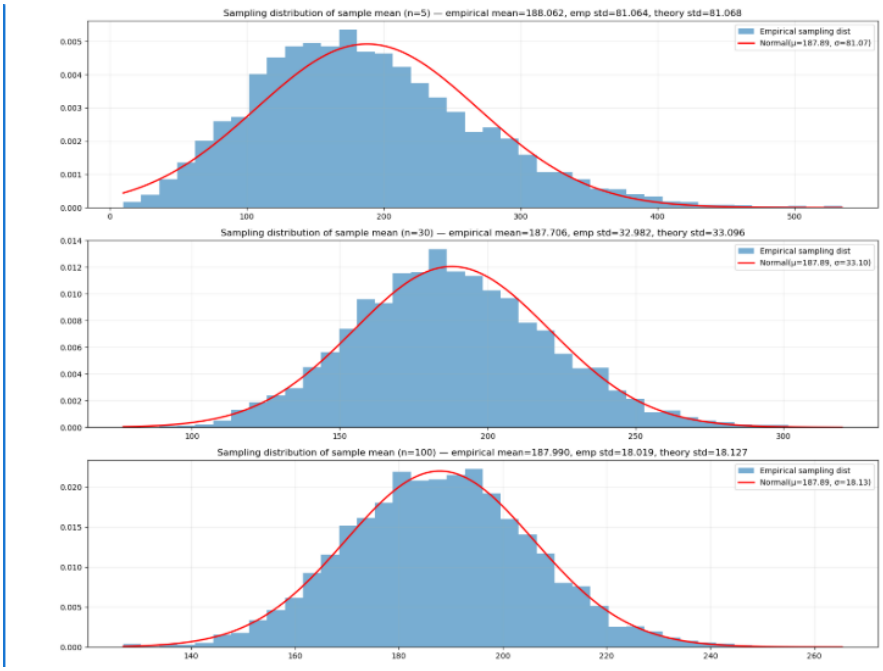
6.1 Law of Large Numbers (LLN) demonstration

- Variable chosen and why it makes sense: Cumulative average is suitable because it tracks the sample mean as samples increases and can easily be compared with population average.
- Experiment: show sample mean as n increases: The cumulative average for different sample sizes were taken and compared with population average.
- Plot and short interpretation:The cumulative average approaches the population average as the number of samples increase.



6.2 Central Limit Theorem (CLT) application

- Sampling procedure (sample size, number of trials, with or without replacement): Sample size [5,30,100], No of Trials [5000], Data samples were drawn with replacement.
- Show distribution of sample means for increasing n:As sample size is increased, the sample mean distribution becomes more symmetric and more concentrated around population mean.
- Plot(s): histogram(s) of sample means and comparison to normal shape:



6.3 Result interpretation

- The sample probability converges to the theoretical (true) probability as the number of samples were increased. Bike rental counts change with different factors and it is important to take large samples to better visualize overall trends.
- The increase in samples brings the distribution to be more symmetric around the population mean. This shows that using a proper sample size is necessary to visualize the overall distribution of the data as a small sample size could capture less frequent values and give them a similar weight to the actual nature of the data.

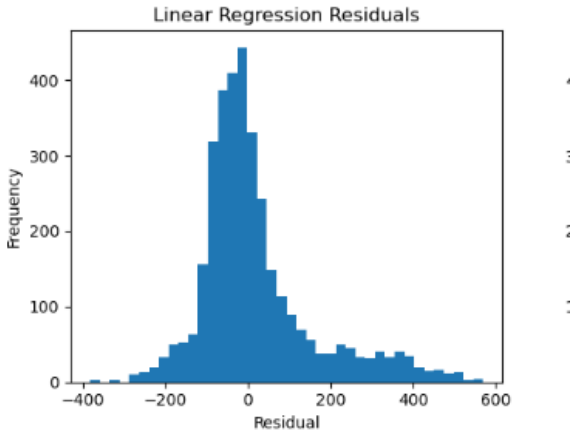
7 Task 5. Regression Analysis

7.1 Linear or Polynomial model selection

- Define target y and predictors X: Target - [Bike rental count], Predictors - [hour of day, Weekend status, season, temperature, wind speed]
- Motivation for linear vs polynomial: Linear models were used to avoid overfitting the data. Linear model is better at showing how each predictor affects the target which is important for analyzing seasonal and weather related factors.
- Any train-test split rationale (time-aware split if relevant):Training was performed on earlier observation and testing on later observations.

7.2 Model fitting and validation

- Fit procedure and preprocessing (scaling, feature selection):Numerical features were standardized via StandardScalar
- Validation method (holdout, time-series split, etc.):A time based split was used to train and test. Earlier data was used to train and later data was used to test.
- Metrics reported (RMSE, MAE, R^2) and why:RMSE(141.390) - [To represent large prediction errors.] MAE(96.23) - [To represent the actual error deviation] R^2(0.589)- [To represent how well the model predicts variation in data.]
- Residual analysis (at least one plot recommended):



7.3 Result interpretation and analysis

- Main effects and practical meaning: Linear regression residuals are centered around zero with no systematic problems but right skewed showing the model fails at high demand hours.
- Failure cases or where model performs poorly: Long right tail shows that the model performs poorly for extreme values.

8 Bonus Tasks

- New dataset bonus (10): state why dataset is new and provide link:
- Q-Q plot with explanation (5):
  - Either for CLT sample means, or regression residuals:
  - Interpretation of deviations from normality:
- Interactive visualizations (up to 10): describe tool used and what interactivity adds:
- Cross-validation in regression (5): method used and how results compare to holdout:
- Additional exploration (up to 20): clearly state extra tasks and value gained:

9 Key Findings and Conclusions

- Main findings from pre-processing and EDA: Rental demand exhibit strong daily and weekly patterns.
- Main findings from probability tasks: Hours occuring in Summer are more likely to exceed the overall median rental demand.
- Main findings from LLN and CLT: Increased samples are required to visualize data as cumulative means converge toward population mean. Mean distributions centered around true mean with increased samples.
- Main findings from regression: Linear regression models capture general demand trends driven by time, season, and recent usage history but perform poorly during extreme demand spikes
- Limitations:Linear models are limited in capturing nonlinear demand dynamics
- What you would do next if you had more time: Incorporate and compare linear models with non-linear models.

10 Reproducibility Notes

- Exact dataset source link and version or download date:(https://archive.ics.uci.edu/dataset/275/bike+sharing
  Downloaded on 19/12/2025
- Key libraries used and versions (optional but recommended):
- How to run the notebook end-to-end:Download the notebook and dataset → open Jupyter → open the notebook → Kernel → Restart & Run All.