Data Science on Bike Rental

Prepared by Mina Roohnavazfar

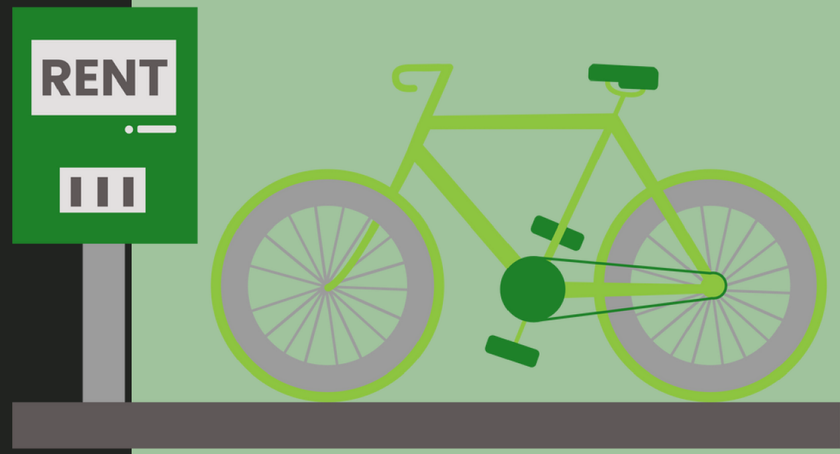
Overview

- POPULARITY OF BIKE SHARING SYSTEMS IN URBAN AREAS
  - SUSTAINABLE AND ENVIRONMENT-FRIENDLY
  - COST-EFFECTIVENESS
  - REDUCED TRAFFIC CONGESTION

- CRITICAL ROLE OF DEMAND FORECASTING IN OPTIMIZING BIKE AVAILABILITY

- PREDICTING THE FUTURE DEMAND PATTERNS

- LEVERAGING HISTORICAL BIKE RENTAL DATA AND APPLYING ADVANCED DATA SCIENCE TECHNIQUES



Data-Driven approach on Bike Rental

- DATA COLLECTION
- DATA PREPROCESSING
- EXPLORATORY DATA ANALYSIS (EDA)
- FEATURE ENGINEERING
- MODEL SELECTION
- MODEL TRAINING AND VALIDATION
- HYPER-PARAMETER TUNING
- DEMAND FORECASTING

Impact of Data-Driven Approach on Bike Rental

ENHANCED BIKE AVAILABILITY

OPTIMIZED RESOURCE ALLOCATION

CITYWIDE TRAFFIC REDUCTION

COST EFFICIENCY

INCREASED REVENUE GENERATION

CUSTOMER SATISFACTION

Dataset Overview

- HISTORICAL BIKE RENTAL DATA FROM THE WASHINGTON D.C. (2011-2012)
- HTTPS://ARCHIVE.ICS.UCI.EDU/DATASET/275/BIKE+SHARING+DATASET
- 17379 RECORDS AND 16 FIELDS
- FEATURES & TARGET VARIABLE

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	instant	data	season	year	month	hour	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
2	1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0	3	13	16
3	2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.8	0	8	32	40
4	3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.8	0	5	27	32
5	4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0	3	10	13
6	5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0	0	1	1
7	6	2011-01-01	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1
8	7	2011-01-01	1	0	1	6	0	6	0	1	0.22	0.2727	0.8	0	2	0	2
9	8	2011-01-01	1	0	1	7	0	6	0	1	0.2	0.2576	0.86	0	1	2	3
10	9	2011-01-01	1	0	1	8	0	6	0	1	0.24	0.2879	0.75	0	1	7	8
11	10	2011-01-01	1	0	1	9	0	6	0	1	0.32	0.3485	0.76	0	8	6	14

Dataset Overview

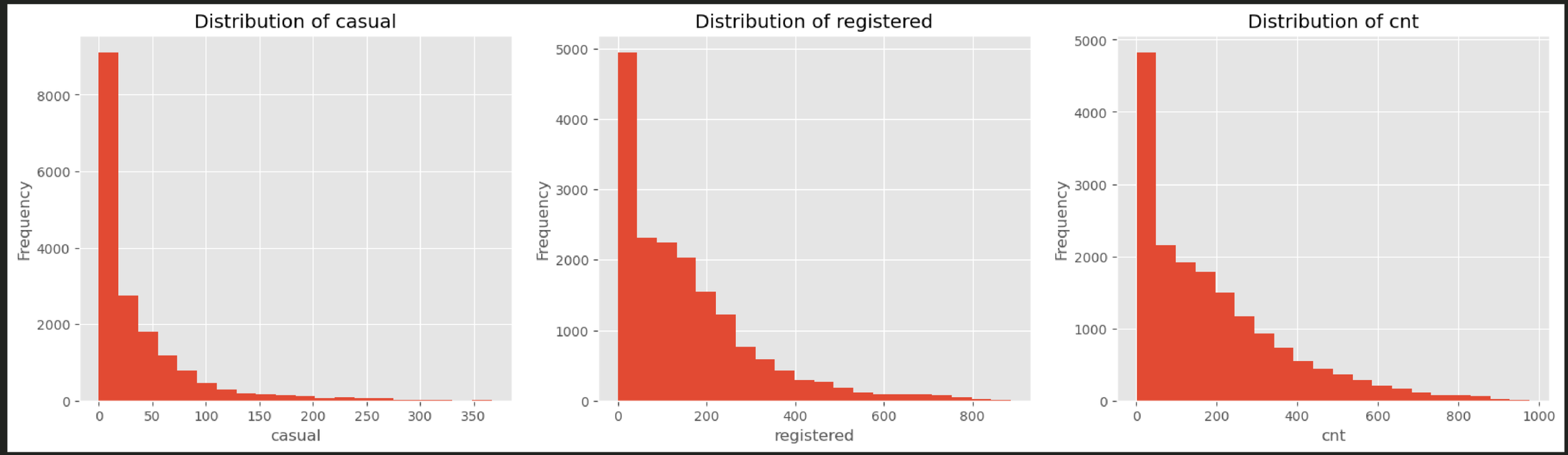
- CONSIDERATION FOR MULTIPLE TARGET VARIABLES:
  - APPROACH 1 : TOTAL COUNT PREDICTION
  - APPROACH 2: SEPARATE USER TYPE PREDICTION

- DATA QUALITY:
  - DATA IS CLEAN IN TERMS OF MISSING VALUES AND DUPLICATE ENTRIES.

- HANDLING CATEGORICAL VARIABLES
- DATA SCALING

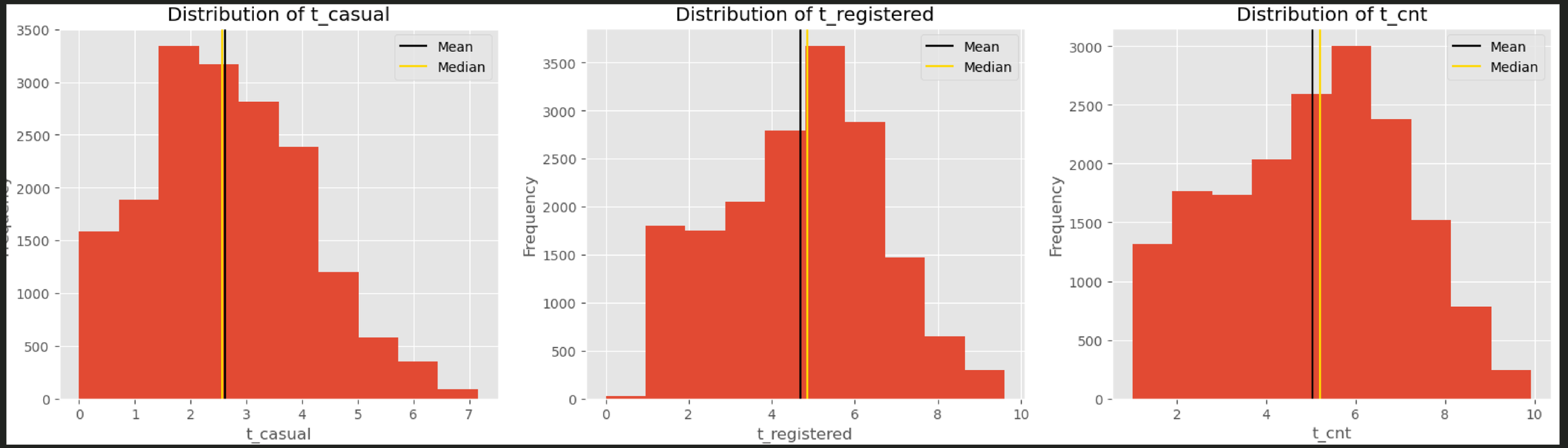
Preliminary EDA Findings

- **Skewness in Distribution of Target Variables**



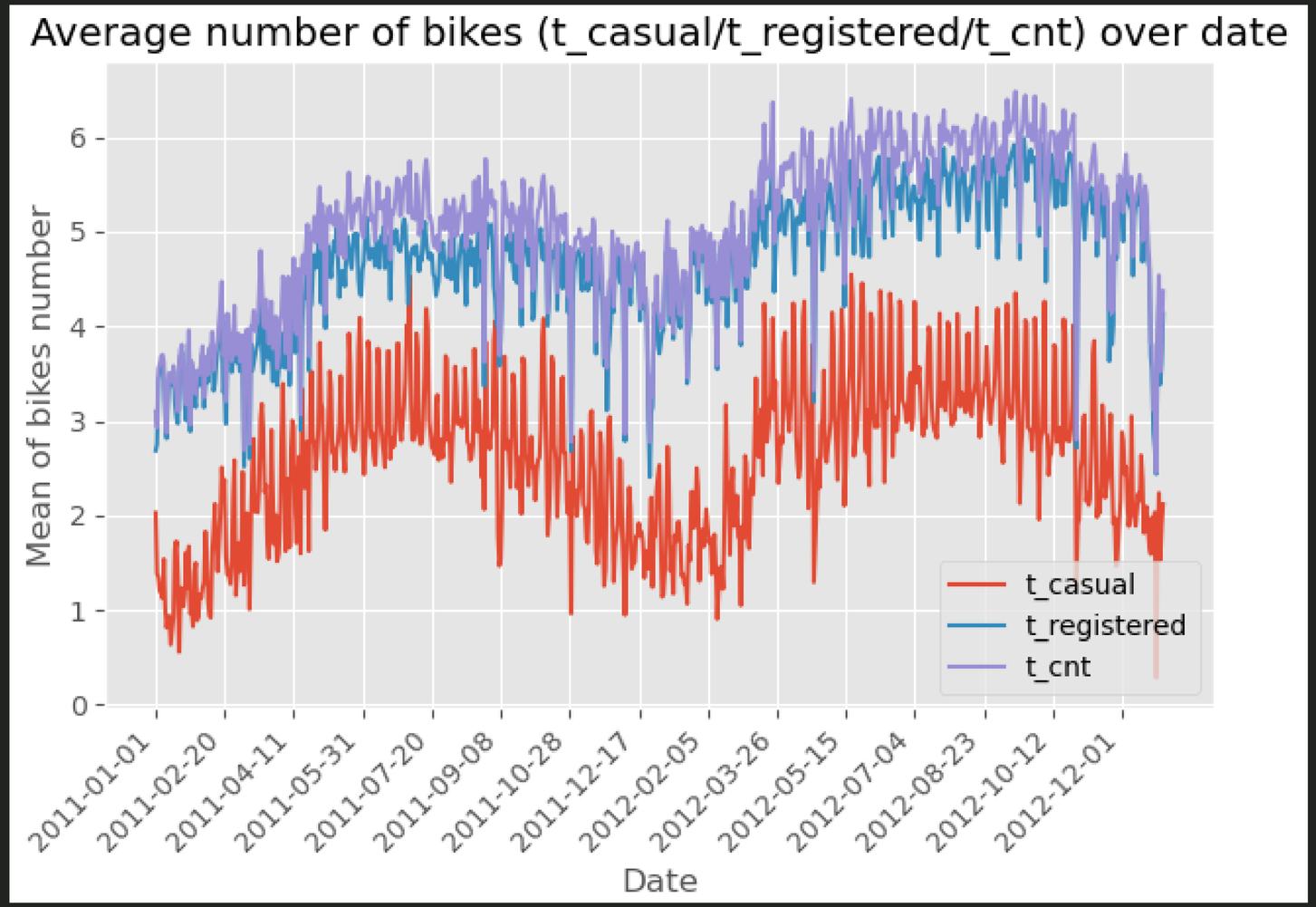
Preliminary EDA Findings

- **Square Root Transformation on Target Variables**



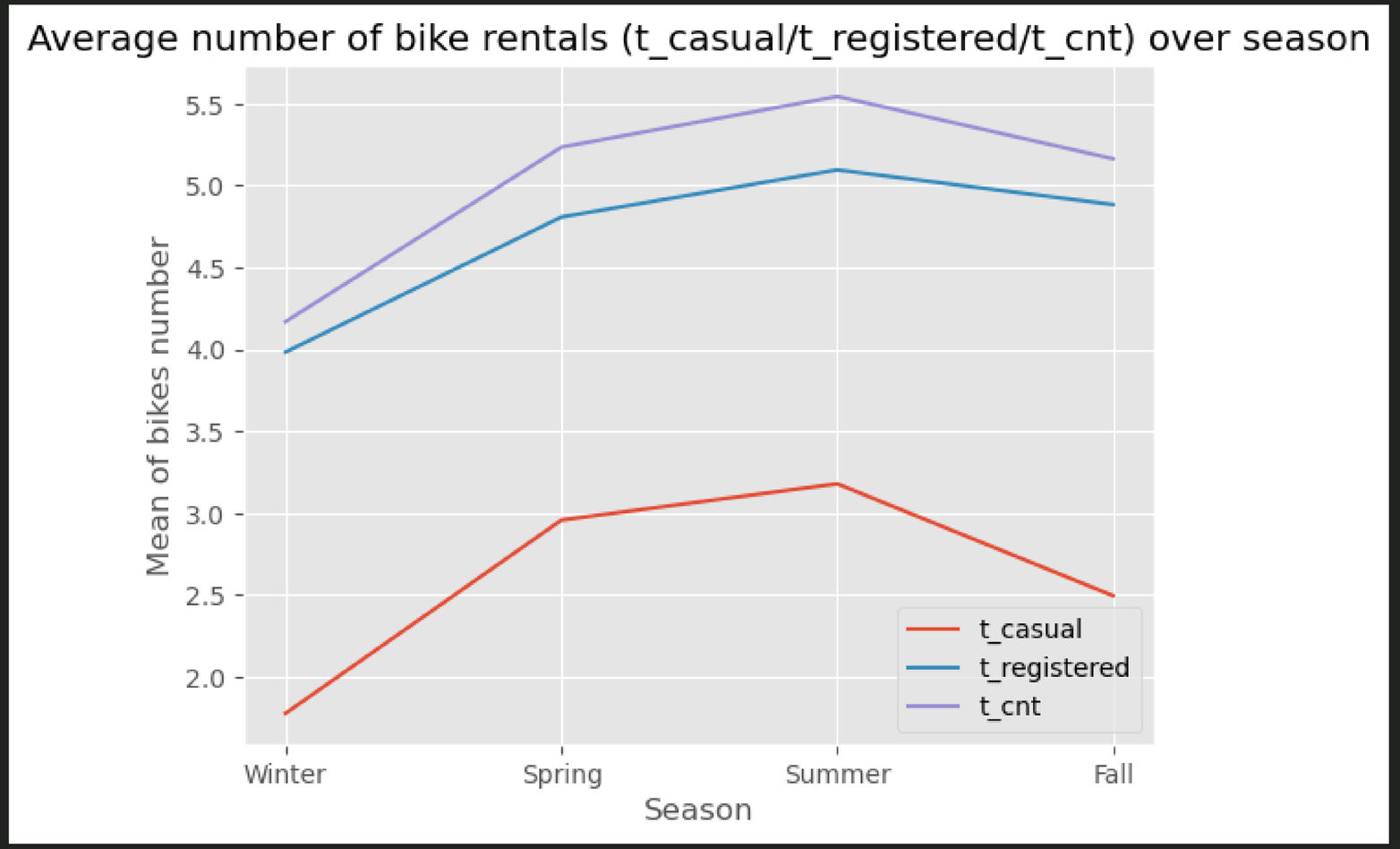
Preliminary EDA Findings

- **Overall Trend**



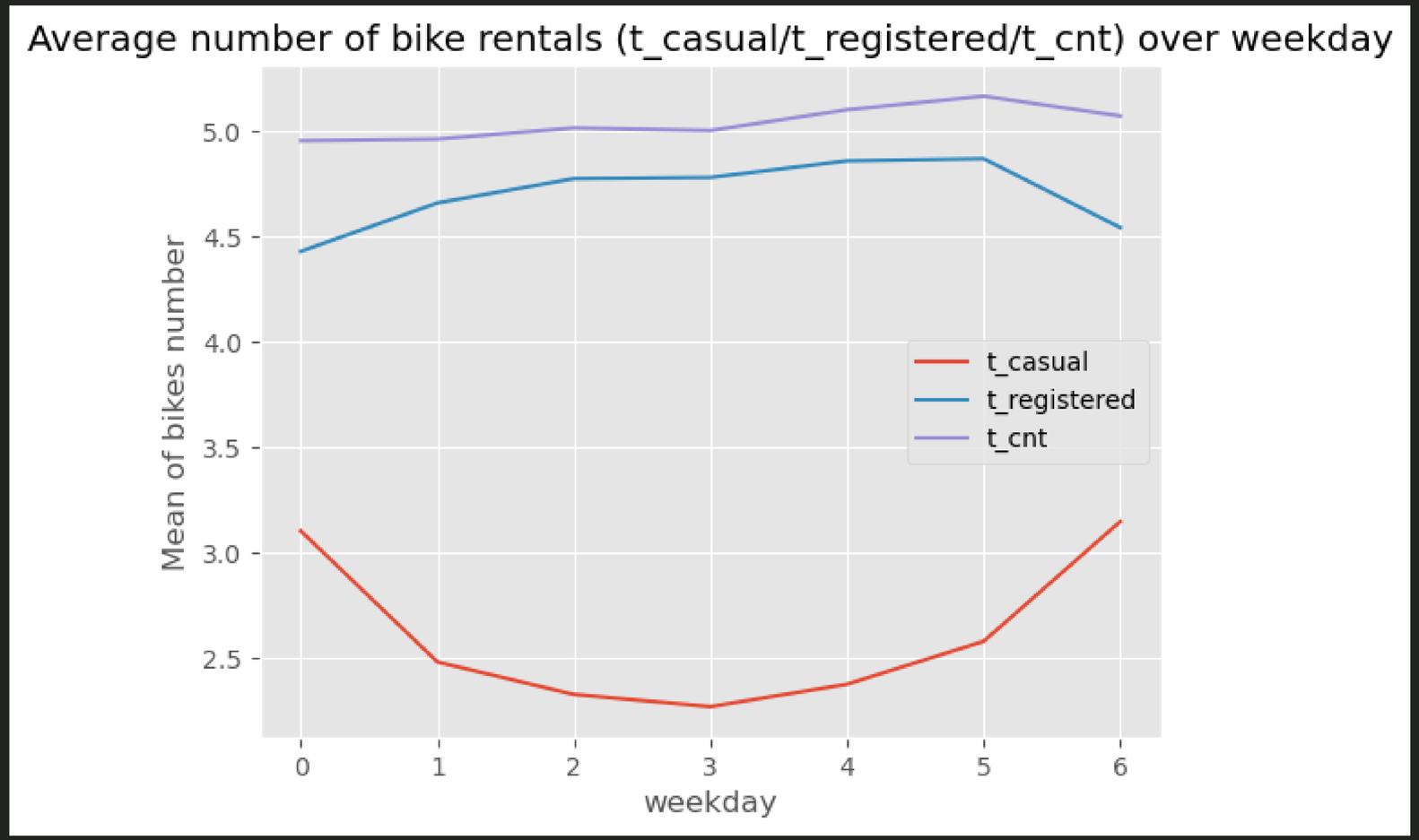
Preliminary EDA Findings

- **Seasonal Variation**



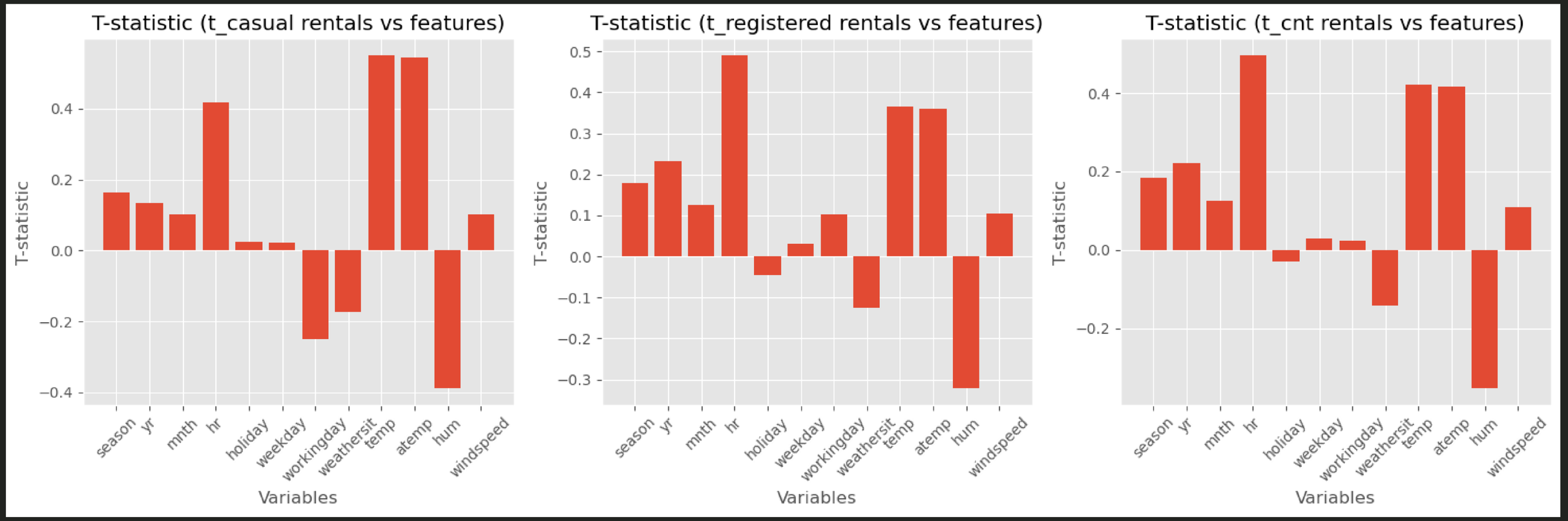
Preliminary EDA Findings

- **Day of the Week**



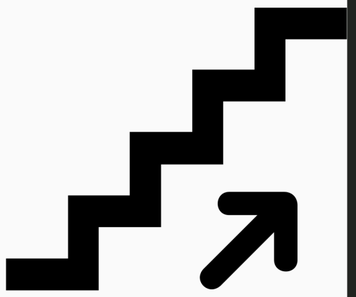
Statistical Analysis

- **Correlation between predictors and Target Variables:**



Future Steps

- BASELINE MODELING:
  - + APPROACH 1:
    - BUILD A LINEAR REGRESSION MODEL TO PREDICT TOTAL COUNT PREDICTION
    - ASSESS THE MODEL'S PERFORMANCE USING EVALUATION METRICS
  - + APPROACH 2: SEPARATE USER TYPE PREDICTION
    - BUILD INDIVIDUAL LINEAR REGRESSION MODELS FOR CASUAL AND REGISTERED USERS
    - COMBINE THE PREDICTIONS OF BOTH MODELS TO OBTAIN THE TOTAL COUNT.
- APPLYING OTHER MODELS: (TIME SERIES MODELS, NEURAL NETWORD, ENSEMBLE METHODS)
- CONDUCT CROSS-VALIDATION
- HYPERPARAMETER TUNING



THANK YOU

FOR YOUR ATTENTION