Columbia University
In the City of New York

BREAKDOWN PREDICTOR PROJECT:

BREAKDOWN TIME PREDICTION FOR DOT FOODS

UP-AND-COMING TEAM

AGENDA

1. Our Team
2. Company Overview
3. Problem Description
4. Data Preparation
5. EDA & Key Message
6. Success Measurement
7. Modeling Progress
8. Results & Recommendation

TEAM

Yichun Fang
Leader

Yanrui Song
Business Specialist

Crystal Qiu
Business Specialist

Yushi Zhao
Data Analyst

Yangchun Yao
Data Analyst

Zhen Yan
Modelling Engineer

Yuxuan Liu
Modelling Engineer

Yujun Xie
Data Visualization Specialist

Kexin Zhang
Data Visualization Specialist

COMPANY OVERVIEW – DOT FOODS

1. One of the **largest** food redistributors in North America.

2. **Sourcing and distributing large volumes of food** from producers to smaller distributors, supermarkets, and other retailers.

3. Offering **a wide range of products**: dry, refrigerated, and frozen.

4. Known for **logistics efficiency**, offering **smaller delivery sizes** and a **vast selection** of goods.

COMPANY OVERVIEW – COMPETITORS

Topco Associates:

providing products and services to its member retailers, boosting their purchasing power and competitive pricing in the grocery sector.

Ingredion:

specializing in ingredient solutions, providing a diverse range of starches, sweeteners, and other food ingredients to manufacturers.

US Foods:

providing diverse food and non-food products to restaurants, emphasizing innovation and customer service.

COMPANY OVERVIEW – KEY CONCEPTS IN BUSINESS

1. **Redistributor:** Buys in bulk, sells in smaller quantities.

2. **Foodservice Distributors:** Supply food to institutions like restaurants and schools.

3. **Supply Chain Optimization:** Improves logistics, reduces costs, and enhances efficiency.

4. **Inventory Turns:** Measures how often inventory is sold and replaced.

5. **Fleet Utilization:** Ensures trucks are fully loaded and routes are efficient.

6. **Cold Chain Logistics:** Manages temperature-sensitive goods for safe transport.

PROBLEM IDENTIFICATION

Current Issue

Varying Unloading Times

Over-allocation of Delivery Time

Fleet Utilization Inefficiencies

Goal

Build a model that more accurately predicts the unloading time.

Data Preparation

Data Preparation

Data Merging

New combined Dataset: 151737 rows, 44 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 151737 entries, 0 to 151736
Data columns (total 44 columns):
 #   Column                     Non-Null Count   Dtype
---  ------                     --------------   -----
 0   Load                       151737 non-null  object
 1   Stop                       151737 non-null  int64
 2   Driver Region              151737 non-null  object
 3   Driver ID                  151737 non-null  object
 4   Driver                     151737 non-null  object
 5   Company ID                 151737 non-null  object
 6   Company                    151737 non-null  object
 7   Address                    151737 non-null  object
 8   City                       151737 non-null  object
 9   State                      151737 non-null  object
 10  Routing Region             151737 non-null  object
 11  Miles From Prev Stop       151737 non-null  int64
 12  Weight                     151732 non-null  float64
 13  Cube                       151732 non-null  float64
 14  Piece Cnt                  151732 non-null  float64
 15  Lines                      151732 non-null  float64
 16  Planned Time               140335 non-null  float64
 17  Print Of DR                151378 non-null  datetime64[ns]
 18  Load Due Out               151723 non-null  datetime64[ns]
 19  DC Departure               151228 non-null  datetime64[ns]
 20  Appt                       151737 non-null  datetime64[ns]
 21  GPS Arrival                151737 non-null  datetime64[ns]
 22  GPS Departure              151737 non-null  datetime64[ns]
 23  Time Diff                  151737 non-null  int64
 24  Per                        151737 non-null  int64
 25  STC                        52913 non-null   object
 26  FCFS                       151717 non-null  float64
 27  Start Time                 151737 non-null  datetime64[ns]
 28  Dock Time                  151737 non-null  object
 29  Dock Time Converted        151737 non-null  float64
 30  Appt Week                  151737 non-null  int64
 31  Trailer Space Utilization  151732 non-null  float64
 32  Appt Day                   151737 non-null  object
 33  DepartBeforeAppt           22524 non-null   object
 34  GPSERROR                   7319 non-null    object
 35  LATE                       12426 non-null   object
 36  Skipped                    151737 non-null  object
 37  Trailer Type               151430 non-null  object
 38  Frt Tmp                    151737 non-null  object
 39  Mid Tmp                    151736 non-null  object
 40  Rear Tmp                   151736 non-null  object
 41  Dry Piece Count            151658 non-null  float64
 42  Refrig Piece Count         151657 non-null  float64
 43  Frozen Piece Count         151658 non-null  float64
dtypes: datetime64[ns](7), float64(11), int64(5), object(21)
memory usage: 50.9+ MB
None
```

Data Preparation

Data Cleaning

~~1/1/23 18:08~~	~~0~~	~~1~~	~~STC~~
~~1/1/23 21:36~~	~~0~~	~~1~~	~~STC~~
1/2/23 3:58	53	1	STC
1/2/23 4:38	0	1	STC
1/2/23 4:46	0	1	STC
1/1/23 13:58	0	1	STC

- **Delete non-null STC value**
  STC: Customer is subject to count, exclude from our analysis

- **Unify NA format**
  Driver ID : “UNKNOWN”
  Driver Region: “NONDRI” → Transform to NULL

- **Replace NA with Median**
  Numeric variable like: Weight, Cube, Piece Cnt, Lines, Trailer Space Utilization

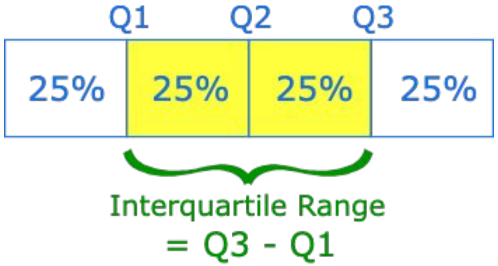
Data Preparation

Data Cleaning

- Fix inconsistencies

Company ID: Set as Null Value

```
['SAGFOR' 'DC01' 'DC16' 'BAKBAK03' 'NORNOR10' 'DC13' 'CHALOT' 'LEBTER'
 'DC19' 'DC41' 'OKLOKL' 'KINGEI' 'STRSTR' 'COSLYO' 'OCATER']
```



- Deal with outliers

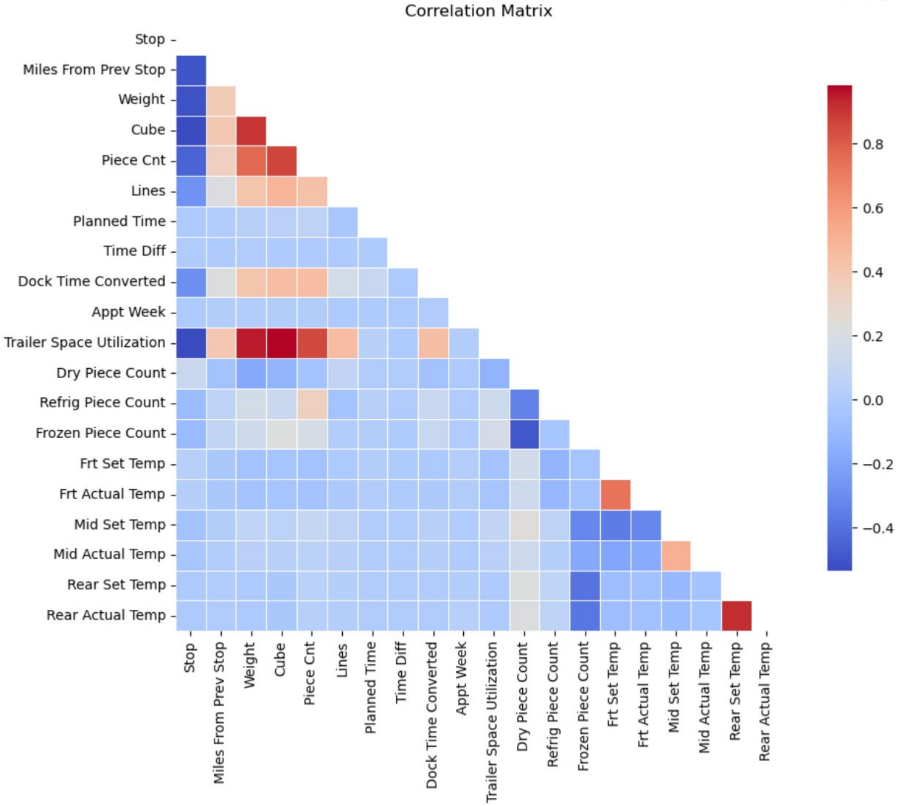
Dock Time Converted:

Drop Outliers with z-score > 3 There are 1737 number of outliers detected by Z-Score.

1 Key Message from Exploratory Data Analysis

Correlation between the Dock Time Converted and others

```
Dock Time Converted           1.000000
Cube                          0.455143
Piece Cnt                     0.453070
Trailer Space Utilization     0.448520
Weight                        0.408301
Miles From Prev Stop          0.225693
Lines                         0.166422
Refrig Piece Count            0.117196
Frozen Piece Count            0.108294
Planned Time                  0.108136
Mid Set Temp                  0.033118
Mid Actual Temp               0.023670
Rear Actual Temp              0.004105
Appt Week                     0.001964
Rear Set Temp                 0.001296
Frt Actual Temp              -0.009698
Time Diff                    -0.012543
Frt Set Temp                 -0.013734
Dry Piece Count              -0.057755
Stop                         -0.292401
Name: Dock Time Converted, dtype: float64
```



1

Key Message from Exploratory Data Analysis

Calculate variance_inflation_factor (vif) for Feature Engineering

```
The VIF of selected Variables:
                    feature         VIF
0                      Cube  115.846582
1                 Piece Cnt   10.088543
2  Trailer Space Utilization  213.822593
3                    Weight   38.233379
4       Miles From Prev Stop    2.021738
5                     Lines    2.353255
6                      Stop    1.537385
```

2

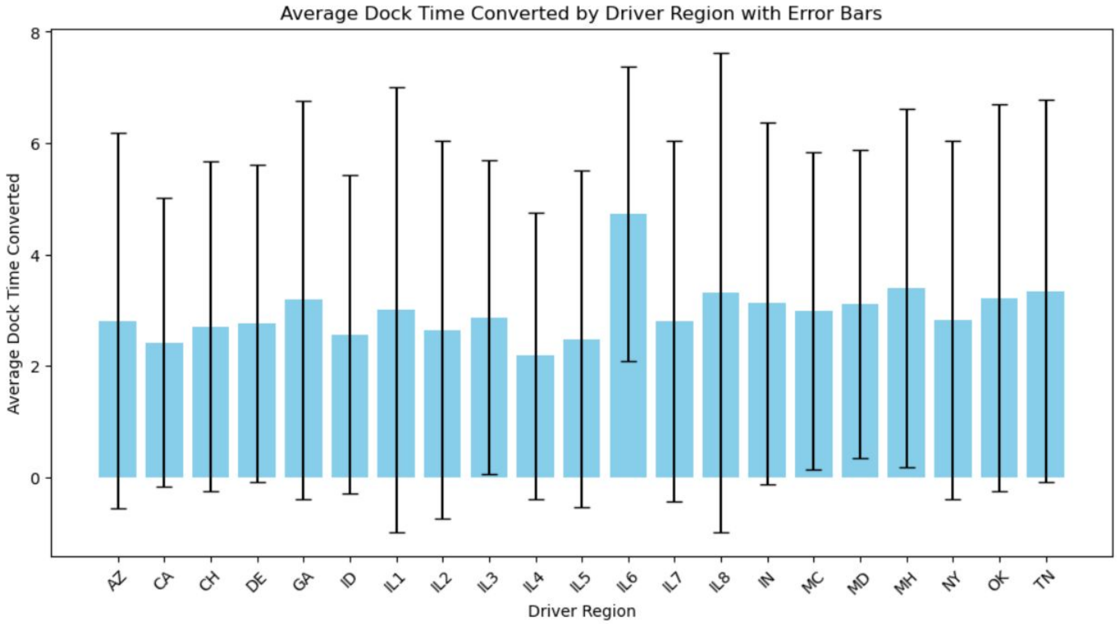
Key Message With Important Categorical Variables

Impact of Driver Region on Dock Time

ANOVA Test Statistic: 45.304369

P-value: 2.9e-178

Regions like **IL6** and **MH** show **higher** dock times, suggesting potential unloading inefficiencies, while **IL4** and **NY** have shorter times, indicating greater efficiency. Large error bars across regions highlight variability, likely due to operational differences or other factors.



Key Message With Important Categorical Variables

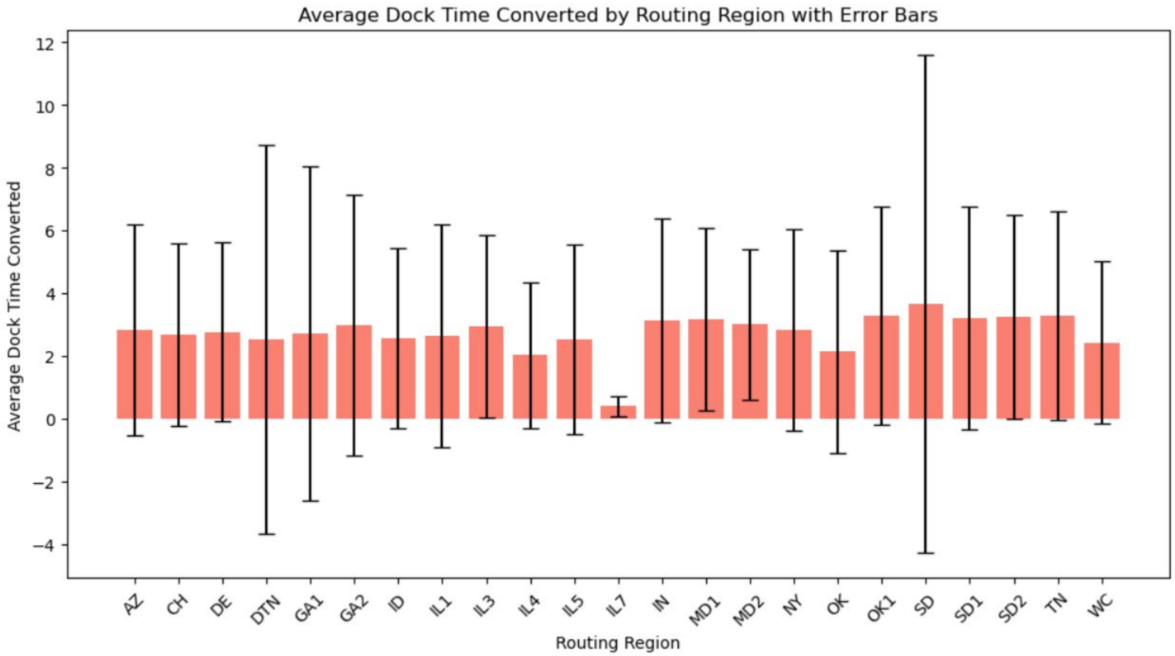
3

Impact of Routine Region on Dock Time

ANOVA Test Statistic: 47.001502

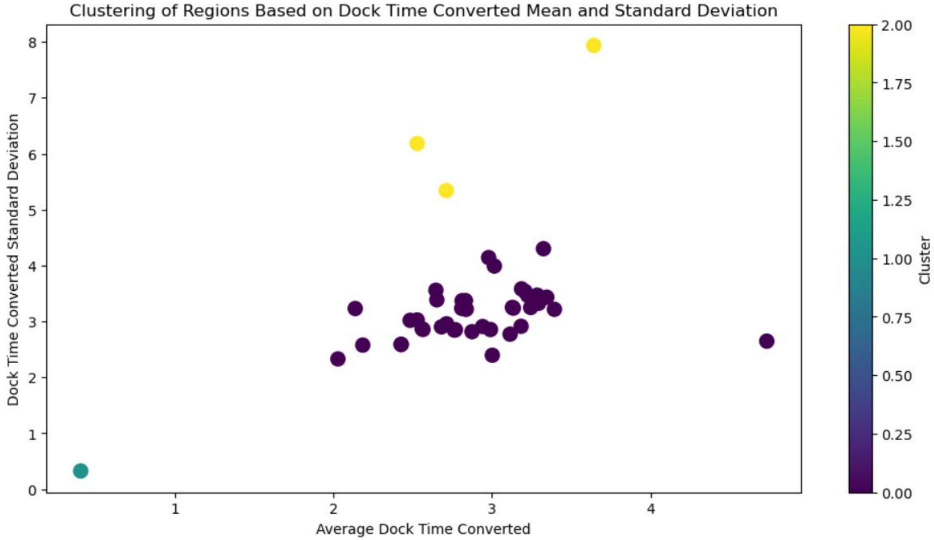
P-value: 1.501668e-203

Regions like **SD** and **DTN** have **higher** dock times, indicating potential unloading inefficiencies, while **IL7** and **TN** show shorter times, suggesting greater efficiency.
Large error bars reveal high variability within regions, likely due to differing local practices or external factors.



3 Key Message With Important Categorical Variables

Clustering of Regions Based on Dock Time Converted Efficiency and Variability



Clustering Method

- **Cluster 0**: Regions with moderate dock times and low variability
- **Cluster 1**: High-efficiency region with low dock time and minimal variability
- **Cluster 2**: Regions with high variability, indicating potential inconsistency in processes

Result:
Routing Region DTN 2
Routing Region GA1 2
Routing Region IL7 1
Routing Region SD 2
The other regions all 0

Conclusion

- Regions in Cluster 2 (e.g., SD, DTN) show high variability and may need process standardization.
- Cluster 1 (IL7) is a model of efficiency, with potential best practices to apply to other regions.
- Most regions fall in Cluster 0, with relatively stable dock time

3

Key Message With Important Categorical Variables

Clustering of Variables to Reduce Complexity & Improve Interpretability

Variable	F-value	P-value
Driver	4.905554	0.0
Company	16.145481	0.0
Address	16.85368	0.0
City	21.518489	0.0

Purpose of Clustering:
Simplifies high-dimensional variables (`Driver`, `Company`, `Address`, `City`) by grouping them into performance tiers based on dock time efficiency.

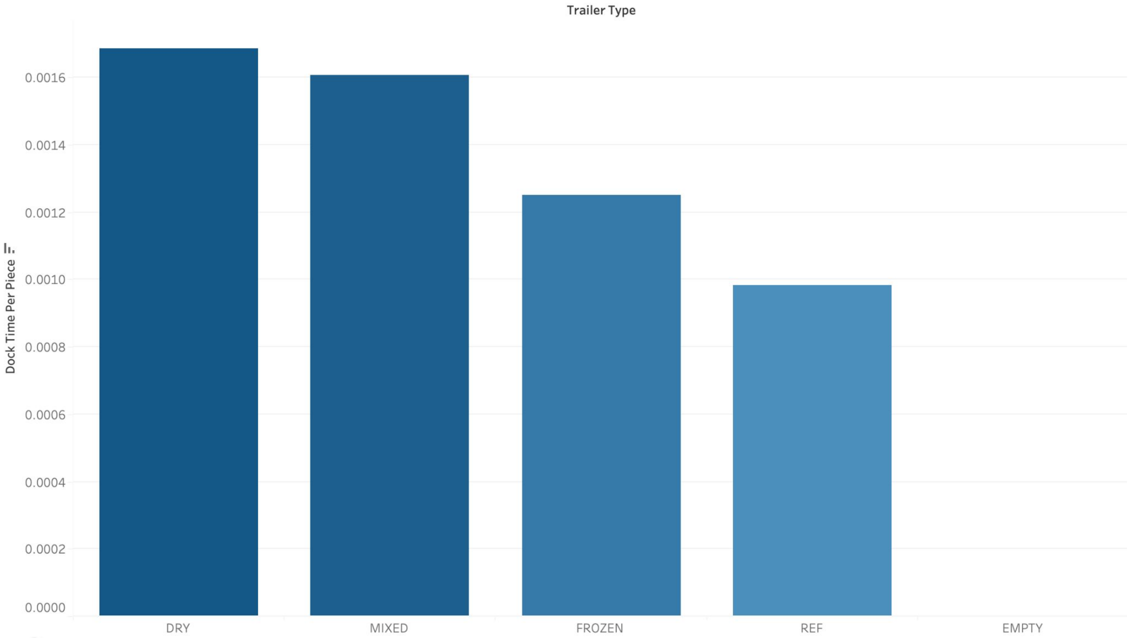
Method: Calculated average dock time for each category and clustered into high, medium, low efficiency groups using K-means.

Result: Reduces model complexity and maintains interpretability, as shown by statistically significant F-values and low P-values for each variable.

4

Key Message from Exploratory Data Analysis

Segment Analysis - Impact of Trailer Type on Dock Time



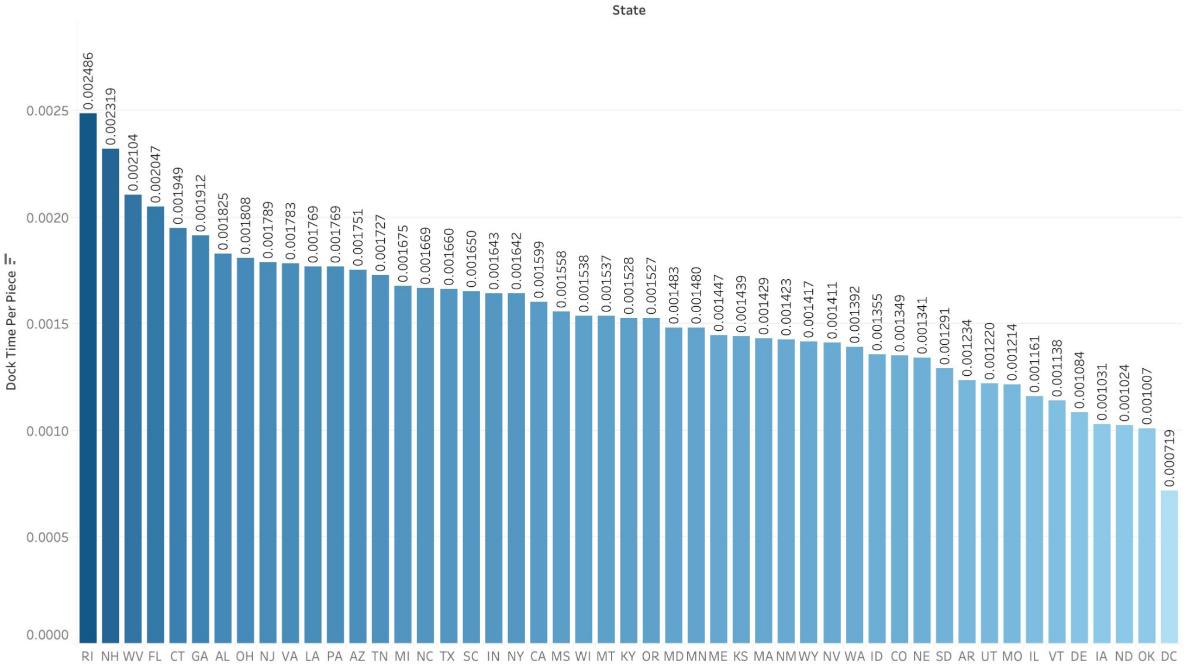
ANOVA Test Statistic: 2.45

P-value: **0.061**

Although **dry trailers** seem to take longer on average and **refrigerated trailers** tend to have shorter dock times, the results do not reach statistical significance.

5 Key Message from Exploratory Data Analysis

Segment Analysis - Impact of State on Dock Time



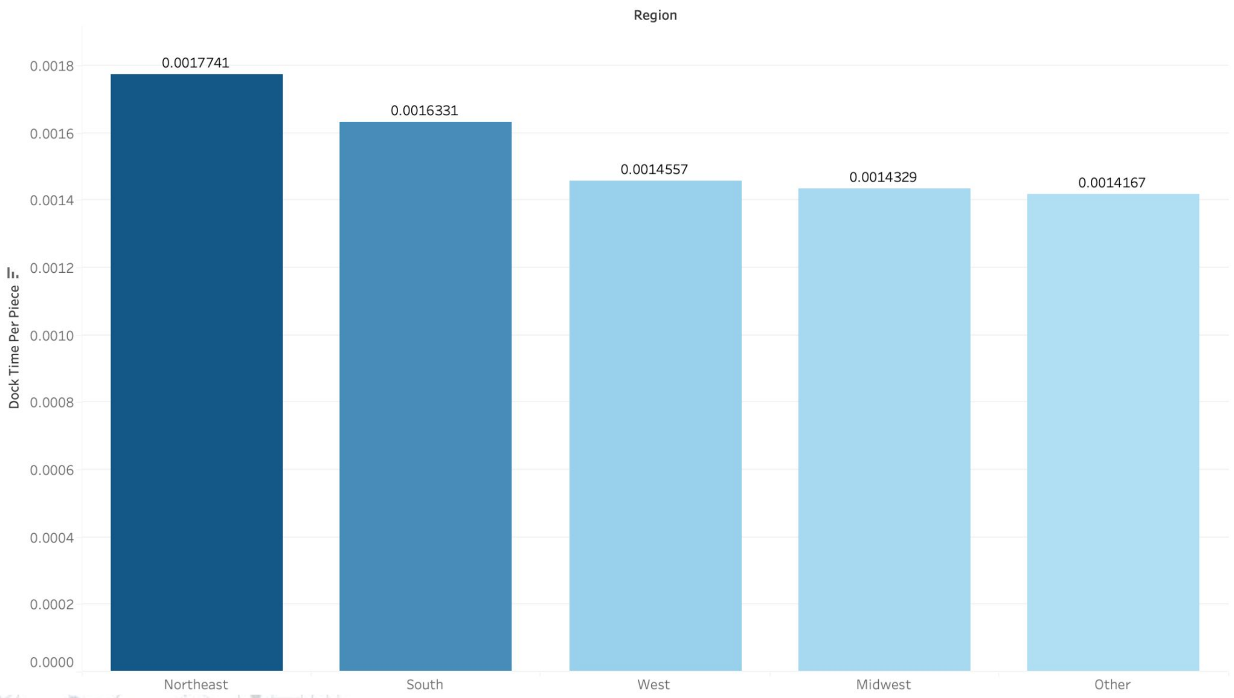
ANOVA Test Statistic: 37.41

P-value: **0.0 (highly significant)**

The differences in dock times across states are **statistically significant**, meaning dock times are not random but influenced by geographical location.

5 Key Message from Exploratory Data Analysis

Segment Analysis - Impact of State on Dock Time

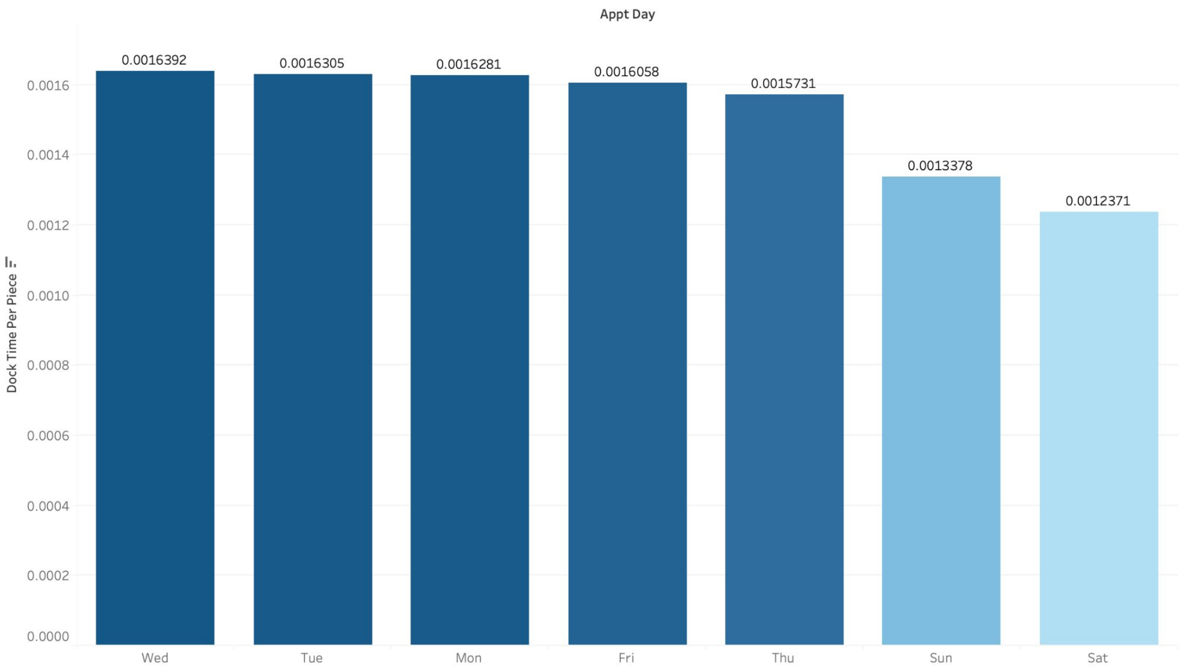


Northeast and **South** regions have the highest dock times, suggesting potential inefficiencies in unloading processes. **Other** region shows the shortest dock times, indicating more efficient unloading operations compared to other regions.

6

Key Message from Exploratory Data Analysis

Segment Analysis - Impact of Appointment Day on Dock Time



ANOVA Test Statistic: 17.33

P-value: **3.95e-20**

Weekends have notably lower dock time per piece compared to weekdays, indicating a faster unloading process on these days. **Wednesday** and **Tuesday** show the highest dock time per piece, suggesting that these days are the most time-intensive for unloading.

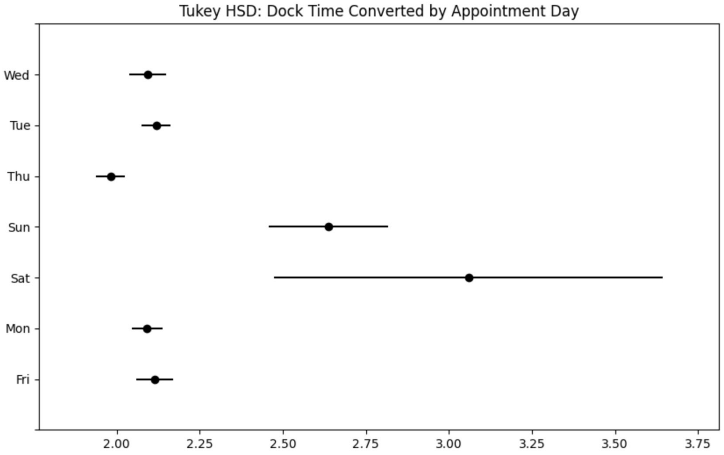
6 Key Message from Exploratory Data Analysis

Tukey's HSD Post-Hoc Test Results for Appointment Day

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Fri	Mon	-0.022	0.994	-0.1181	0.0741	False
Fri	Sat	0.9451	0.0004	0.2956	1.5946	True
Fri	Sun	0.5237	0.0	0.2761	0.7714	True
Fri	Thu	-0.1323	0.0006	-0.2258	-0.0387	True
Fri	Tue	0.0049	1.0	-0.0867	0.0964	False
Fri	Wed	-0.0208	0.9977	-0.1289	0.0873	False
Mon	Sat	0.9671	0.0002	0.3194	1.6148	True
Mon	Sun	0.5457	0.0	0.3028	0.7887	True
Mon	Thu	-0.1103	0.001	-0.1906	-0.03	True
Mon	Tue	0.0269	0.9506	-0.0511	0.1048	False
Mon	Wed	0.0012	1.0	-0.0957	0.0981	False
Sat	Sun	-0.4214	0.5417	-1.1081	0.2654	False
Sat	Thu	-1.0774	0.0	-1.7247	-0.43	True
Sat	Tue	-0.9402	0.0004	-1.5873	-0.2932	True
Sat	Wed	-0.9659	0.0002	-1.6155	-0.3163	True
Sun	Thu	-0.656	0.0	-0.898	-0.414	True
Sun	Tue	-0.5189	0.0	-0.7601	-0.2776	True
Sun	Wed	-0.5445	0.0	-0.7925	-0.2966	True
Thu	Tue	0.1371	0.0	0.0623	0.212	True
Thu	Wed	0.1115	0.009	0.0171	0.2059	True
Tue	Wed	-0.0257	0.9831	-0.1181	0.0667	False

Tukey's HSD identifies that **Saturday** has significantly longer dock times than **Thursday** and **Monday**, while other weekdays show no significant differences.



Model Selection

—— Linear Regression & Random Forest & XGBoost

Model Selection

1. Linear Regression:

- **Assumes linear relationships** between features and dock time.
- Focuses on core variables such as:
  - **Number of items** (Piece Count)
  - **Trailer space utilization**
  - **Miles from previous stop**
- Provides a **baseline understanding** of dock time influences.

2. Random Forest:

- **Handles non-linear relationships** and **complex interactions**.
- Includes additional variables:
  - **Trailer type**, **regional factors**
- Offers **feature importance** to highlight key variables influencing dock time.

3. XGBoost:

- **Ensemble model** capturing complex, nuanced interactions.
- Effective in handling:
  - **Missing values**
  - **Non-linear effects** (e.g., driver region, trailer conditions)
- Expected to provide the **most accurate predictions** for dock time.

Success Measurement
Quantitative Performance Metrics

Root Mean Squared Error (RMSE)

- The **square root of the average squared differences** between the predicted and actual breakdown times
- Target: Less than **15** minutes

R^2 (Coefficient of Determination)

- The **proportion of the variation** in actual breakdown times that can be explained by the model's predictions

Adjusted R^2

- Adjusted R^2 modifies R^2 to account for the number of predictors in the model. It provides a more accurate measure of the goodness of fit
-

Data Modeling Preprocessing

Variable Selection (VIF): Selected numeric variables (Stop, Piece Cnt, Lines, Time Diff) with low VIF to reduce multicollinearity.

One-hot Encoding: Driver Region cluster, Routing Region cluster, state cluster(region), Trailer type, appointment day, driver cluster, city cluster , company cluster , address cluster

Log Transformation: Applied log transformation to all numeric variables and the target Dock Time to standardize, making the data less skewed, improve model performance, adding 1 to each value to handle zeros and negatives, preserving proportion.

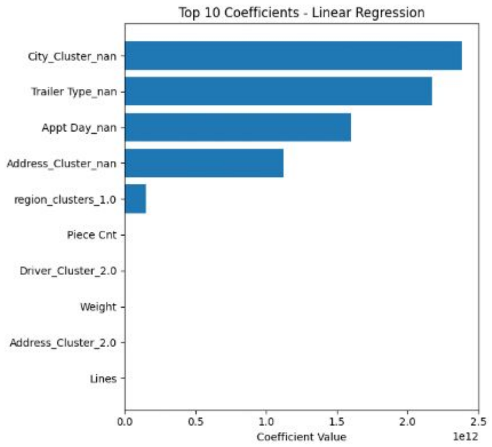
Linear Regression

To Dock Time Converted

Model	RMSE	R^2	Adjusted R^2
Linear Regression	2.64	0.31	0.31
Log-Transformation	2.68	0.29	0.29

Top 10 Features by Linear Regression Coefficients:

Feature	Linear_Coefficient
City_Cluster_nan	2.382596e+12
Trailer Type_nan	2.172414e+12
Appt Day_nan	1.597146e+12
Address_Cluster_nan	1.121465e+12
region_clusters_1.0	1.491698e+11
Piece Cnt	3.874374e-01
Driver_Cluster_2.0	2.908499e-01
Weight	2.355533e-01
Address_Cluster_2.0	2.203473e-01
Lines	1.691268e-01

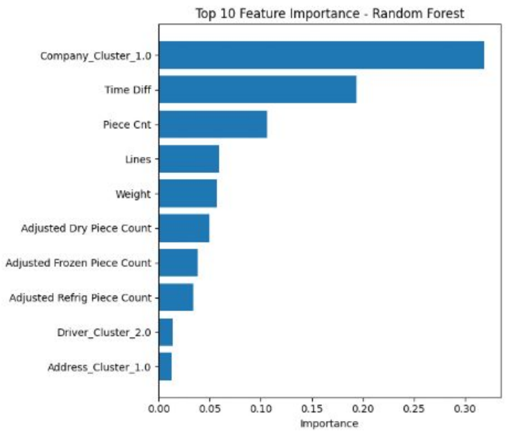


Random Forest

To Dock Time Converted

Model	RMSE	R^2	Adjusted R^2
Random Forest	2.27	0.49	0.49
Log-Transformation	2.37	0.45	0.44

```
Top 10 Features by Random Forest Importance:
                       Feature  RF_Importance
             Company_Cluster_1.0       0.319062
                       Time Diff       0.193999
                       Piece Cnt       0.106439
                           Lines       0.059339
                          Weight       0.057201
         Adjusted Dry Piece Count       0.049612
      Adjusted Frozen Piece Count       0.037984
      Adjusted Refrig Piece Count       0.034180
             Driver_Cluster_2.0       0.013702
            Address_Cluster_1.0       0.012913
```

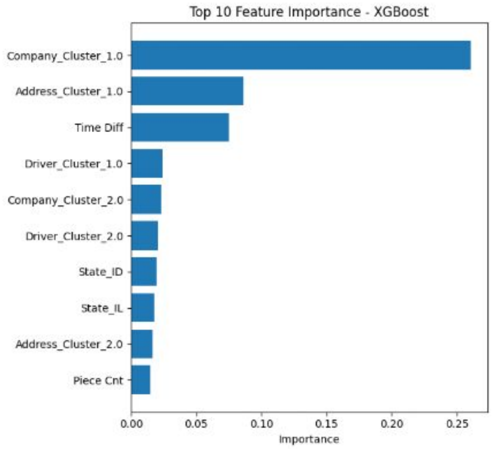


XG Boost

To Dock Time Converted

Model	RMSE	R^2	Adjusted R^2
XG Boost	2.25	0.50	0.50
Log-Transformation	2.33	0.46	0.46

```
Top 10 Features by XGB Importance:
              Feature  XGB_Importance
  Company_Cluster_1.0        0.260582
  Address_Cluster_1.0        0.086366
            Time Diff        0.075248
   Driver_Cluster_1.0        0.024038
  Company_Cluster_2.0        0.023080
   Driver_Cluster_2.0        0.020524
             State_ID        0.019725
             State_IL        0.018002
  Address_Cluster_2.0        0.016617
            Piece Cnt        0.014667
```



Model Selection

Model	RMSE	R^2	Adjusted R^2
Linear Regression	2.64	0.31	0.31
Random Forest	2.27	0.49	0.49
XG Boost	2.25	0.50	0.50

Model Comparison:

Linear Regression

Strengths:Easy to understand, interpretable, no tuning required

Random Forest Regressor

Strengths: Robust to noise, suitable for smaller datasets or simpler patterns

XGBoost Regressor

Strengths: Handles complex patterns, reduces overfitting, ideal for large datasets

- ## Conclusion:

According to the previous analysis of the three models, the model where the independent variables are not log-transformed performs better, so we no longer log-transformed. Overall, XG Boost works best among the three models.

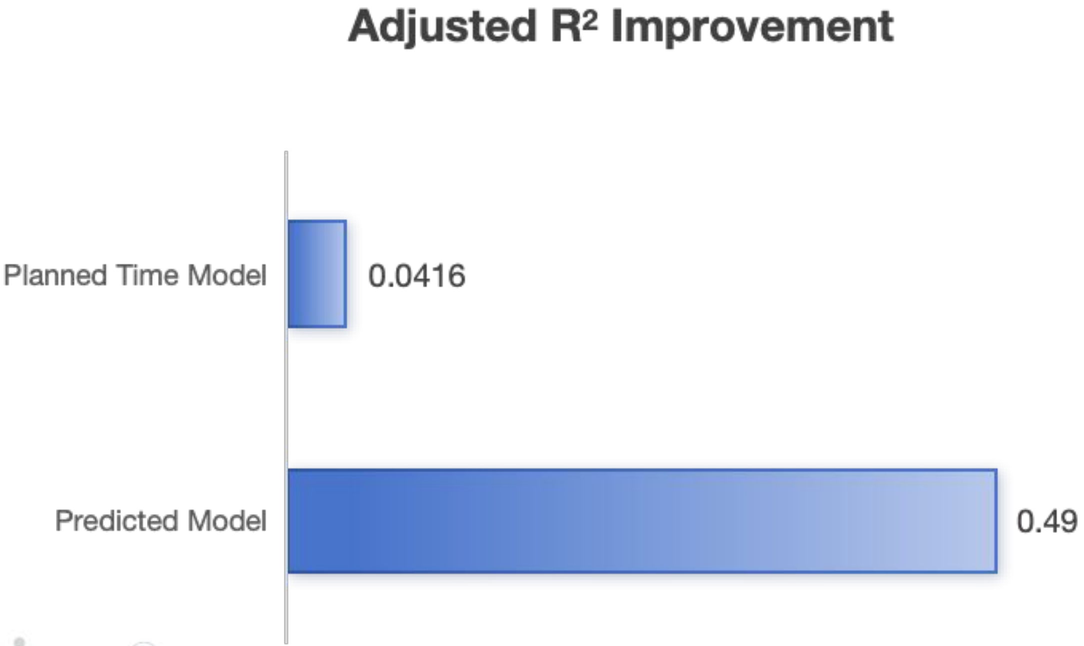
Final Model

Model	RMSE	R^2	Adjusted R^2
XG Boost	2.25	0.50	0.50

Numeric Variables Used: Stop, Piece Cnt, Lines, Time Diff

Categorical Variables Used: Driver Region, Routing Region, State, Trailer Type, Appointment Day, Driver, City, Company, Address

Model Effectiveness



· **Better fleet utilization**

More accurate scheduling means trucks are used more efficiently.

· **Reduced operational costs**

Minimizing unnecessary delays and improving resource allocation.

· **Increased customer satisfaction**

Timely and reliable deliveries contribute to stronger relationships with customers.

Thank you!