DATA ANALYTICS 1 (ADTA 5130) – FINAL PROJECT (GROUP 25)

Course Name: Data Analytics 1

Course Code: ADTA 5130

Submitted by: Team 25

1) Vishwa Tej Reddy Chitla (11645954)

2) Lakshmi Prasanna Kommineni (11648973)

3) Zakiya AI Jabri (11727401)

OUTLINE

Research Questions & Expected Outcomes

In-depth analysis and comprehension of the complex dynamics included in hotel reservation data from 2015 to 2017 is the main goal of this research project. This research endeavors to identify the complex variables affecting hotel reservations and their consequent effects on critical facets of the hospitality sector using a thorough analysis and implementation of many statistical models. This analysis aims to identify underlying patterns, trends, and dependencies in the data by examining the large dataset, which contains a variety of characteristics such as visitor preferences to booking specifications.

Research Question 1:

Research Question: What is the impact of arrival month on Average Daily Rate (ADR) across the years 2015-2017?

Null Hypothesis: H0: There is no significant difference in ADR across different months/seasons between 2015 and 2017.

Alternate Hypothesis: HA: There is a significant difference in ADR across different months/seasons between 2015 and 2017.

Expected Outcome: It is hypothesized that there will be variations in ADR based on arrival months, with specific months potentially demonstrating higher ADR compared to others.

Research Question 2:

Research Question: How are hotel types, stays in weekend nights, and stays in weeknights connected in hotel reservations from 2015 to 2017?

Null Hypothesis: H0: There is no relationship between hotel types, weekend stays, and weeknight stays.

Alternate Hypothesis: HA: There is a relationship between hotel types, weekend stays, and weeknight stays.

Expected Outcome: It is anticipated that certain hotel types will attract more weekend or weeknight stays, impacting the overall concentration of customers in specific hotel segments.

Research Question 3:

Research Question: What influence do hotel types and the number of booking changes exert on Average Daily Rate (ADR)?

Null Hypothesis: H0: Hotel types and booking changes have no substantial effect on ADR.

Alternate Hypothesis: HA: Hotel types and booking changes have a significant effect on ADR.

Expected Outcome: The hypothesis posits that variations in hotel types and booking changes will significantly affect ADR, with certain types or changes contributing more to ADR fluctuations.

Research Question 4:

Research Question: Do lead time and days on the waiting list impact hotel reservation cancellations?

Null Hypothesis: H0: Lead time and days on the waiting list do not correlate with hotel reservation cancellations.

Alternate Hypothesis: HA: Lead time and days on the waiting list are correlated with hotel reservation cancellations.

Expected Outcome: It is expected that longer lead times and increased days on the waiting list will correspond to a higher probability of hotel reservation cancellations.

Research Question 5:

Research Question: How does the choice of agent affect Average Daily Rate (ADR) in hotels?

Null Hypothesis: H0: The choice of agent does not significantly influence ADR in hotels.

Alternate Hypothesis: HA: The choice of agent significantly influences ADR in hotels.

Expected Outcome: The hypothesis suggests that different agents will have varying impacts on ADR, potentially allowing hotels to optimize agent choices for increased ADR.

Introduction

Hotels handle a great amount of data on guest reservations. This study examines data from 2015 to 2017 in order to figure out why and how people make hotel reservations. We're investigating factors like arrival times, hotel preferences, and any modifications made to reservations. By doing this, we hope to determine the ways in which these factors impact daily hotel revenue and the

number of cancellations. Our intention is to make better judgments and provide customers with an improved stay experience for hotels by using this information.

The principal dataset employed for this study is the 'hotel_bookings' dataset, which includes 32 unique variables and 119,391 observations of hotel booking data from 2015 to 2017. The variables in this dataset are of a variety of forms, representing different facets of hotel reservations, such as binary, category, and numeric data. Excel functions were used to extract summary statistics for numeric variables including lead time, stays in weekend nights, stays in weeknights, adults, and so on in order to understand the data distribution and core patterns. These statistics included mean, median, minimum, and maximum. Additionally, a comprehensive analysis was carried out to detect any missing values in all columns. The dataset's integrity was improved for further analysis by replacing missing values with the mean of non-null values within the corresponding columns to guarantee data completeness.

APPROACH

Data Collection and Preprocessing:

1. Gathered the 'hotel_bookings' dataset and conducted data cleaning, addressing missing values, outliers, and data type inconsistencies.
2. Formatted date variables and ensured data integrity for data analysis.

Exploratory Data Analysis (EDA):

1. Perform descriptive statistics on numerical variables (e.g., lead time, weekend night stays) to figure out their distributions and central tendencies.
2. Utilized histograms, bar charts, and pie charts to visualize categorical data (e.g., customer_type) in order to recognize patterns and trends.
3. Investigated correlations between variables.

Evaluation of Statistics:

1. The impact of arrival month on the Average Daily Rate (ADR) and the correlation between hotel categories and weekend/weeknight stays were examined using ANOVA (Analysis of Variance) models.
2. To determine how hotel type, booking modifications, and other factors affected ADR, linear regression models were used.
3. The factors impacting hotel reservation cancellations were identified by the application of linear regression.
4. To comprehend the influence of the agent selection on ADR, Linear Regression models were assessed.

Assessment and Interpretation of the Model:

1. Assessed model performance using relevant metrics, including significance levels, F-statistics, and R-squared.
2. Interpreted model summaries, significance tests, and coefficients to provide practical management advice for hotels.

IMPLEMENTATION

ONE-WAY ANOVA MODEL

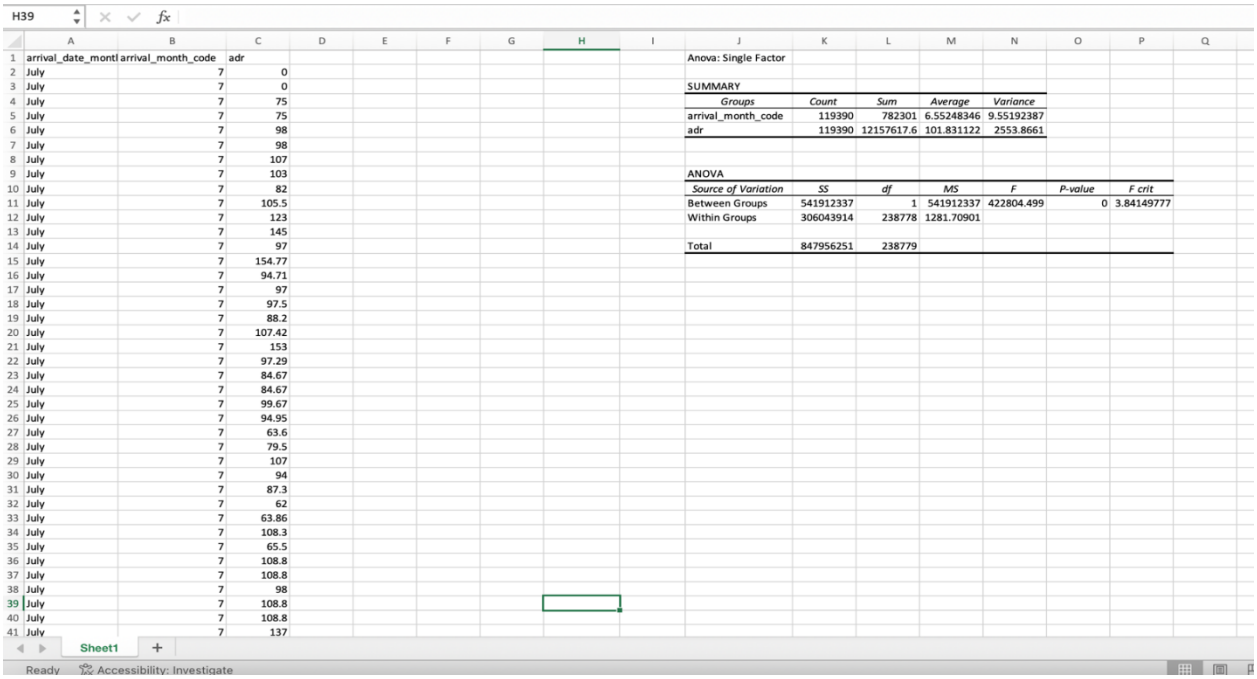
Conducted ANOVA to assess the influence of "arrival_month" on "ADR" (Average Daily Rate) during the period 2015-2017.

Data Preparation:

1. 'ADR' (Average Daily Rate) and 'arrival_date_month' are two of the columns that should be included in the 'hotel_bookings' dataset.
2. To guarantee accurate and comprehensive data for the analysis, filter and clean the dataset.
3. Assign a code to each month to convert the categorical "arrival_date_month" column to a numerical column.

ANOVA Analysis in Excel:

1. Open Excel and sort the information into columns, one for 'ADR' values and another for 'arrival_date_month' categories.
2. Select 'Data Analysis' from the 'Data' tab.
3. Select 'ANOVA: Single Factor' from the analysis tools available and enter the 'ADR' column as the 'Input Range' and the 'arrival_date_month' column as the 'Factor'.
4. To run the ANOVA analysis, click the 'OK' button.



arrival_date_month	arrival_month_code	adr
July	7	0
July	7	0
July	7	75
July	7	75
July	7	98
July	7	98
July	7	107
July	7	103
July	7	82
July	7	105.5
July	7	123
July	7	145
July	7	97
July	7	154.77
July	7	94.71
July	7	97
July	7	97.5
July	7	88.2
July	7	107.42
July	7	153
July	7	97.29
July	7	84.67
July	7	84.67
July	7	99.67
July	7	94.95
July	7	63.6
July	7	79.5
July	7	107
July	7	94
July	7	87.3
July	7	62
July	7	63.86
July	7	108.3
July	7	65.5
July	7	108.8
July	7	108.8
July	7	98
July	7	108.8
July	7	108.8
July	7	137

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
arrival_month_code	119390	782301	6.55248346	9.55192387
adr	119390	12157617.6	101.831122	2553.8661

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	541912337	1	541912337	422804.499	0	3.84149777
Within Groups	306043914	238778	1281.70901			
Total	847956251	238779				

TWO-WAY ANOVA MODEL

Conducted analysis comparing average stays_in_weekend_nights and stays_in_week_nights across different hotel types (e.g., Resort Hotel, City Hotel) over the period 2015-2017.

Data Preparation:

1. Make sure the "hotel_bookings" dataset has the columns "hotel," "stays_in_weeknights," and "stays_in_weekend_nights" that are required.
2. To ensure accurate analysis, filter and clean the dataset to eliminate any null or irrelevant entries.

ANOVA Analysis in Excel:

1. Organize the information in Excel by creating columns for “hotel”, “stays_in_weekend_nights”, and “stays_in_weeknights”.
2. "Select 'Data Analysis' from the 'Data' tab. Select 'ANOVA: Two-Factor Without Replication' and enter 'hotel' as the 'Factor' and 'stays_in_weekend_nights' or'stays_in_week_nights' as the 'Input Range'.
3. Execute the analysis and go over the Excel-generated ANOVA output.

Hotel	stays_in_we	stays_in_weekend_nights				hotel			Anova: Two-Factor Without Replication				
1	0	0				Resort Hotel							
1	0	0				Resort Hotel			*SUMMARY*	*Count*	*Sum*	*Average*	*Variance*
1	1	0				Resort Hotel			1	2	0	0	0
1	1	0				Resort Hotel			1	2	0	0	0
1	2	0				Resort Hotel			1	2	1	0.5	0.5
1	2	0				Resort Hotel			1	2	1	0.5	0.5
1	2	0				Resort Hotel			1	2	2	1	2
1	2	0				Resort Hotel			1	2	2	1	2
1	3	0				Resort Hotel			1	2	2	1	2
1	3	0				Resort Hotel			1	2	2	1	2
1	4	0				Resort Hotel			1	2	3	1.5	4.5
1	4	0				Resort Hotel			1	2	3	1.5	4.5
1	4	0				Resort Hotel			1	2	4	2	8
1	4	0				Resort Hotel			1	2	4	2	8
1	4	0				Resort Hotel			1	2	4	2	8
1	4	0				Resort Hotel			1	2	4	2	8
1	4	0				Resort Hotel			1	2	4	2	8
1	1	0				Resort Hotel			1	2	4	2	8
1	1	0				Resort Hotel			1	2	4	2	8
1	4	0				Resort Hotel			1	2	1	0.5	0.5
1	4	1				Resort Hotel			1	2	1	0.5	0.5
1	4	2				Resort Hotel			1	2	4	2	8
1	4	2				Resort Hotel			1	2	5	2.5	4.5
1	4	2				Resort Hotel			1	2	6	3	2
1	5	2				Resort Hotel			1	2	6	3	2
1	5	2				Resort Hotel			1	2	6	3	2
1	5	2				Resort Hotel			1	2	7	3.5	4.5
1	5	2				Resort Hotel			1	2	7	3.5	4.5
1	5	2				Resort Hotel			1	2	7	3.5	4.5
1	5	2				Resort Hotel			1	2	7	3.5	4.5
1	10	4				Resort Hotel			1	2	7	3.5	4.5
1	11	4				Resort Hotel			1	2	7	3.5	4.5
1	8	2				Resort Hotel			1	2	14	7	18
1	4	2				Resort Hotel			1	2	15	7.5	24.5
1	3	1				Resort Hotel			1	2	10	5	18
1	3	1				Resort Hotel			1	2	6	3	2
1	3	1				Resort Hotel			1	2	4	2	2
1	3	1				Resort Hotel			1	2	4	2	2
1	3	1				Resort Hotel			1	2	4	2	2
1	3	2				Resort Hotel			1	2	4	2	2

LINEAR REGRESSION MODEL

Constructed a Linear Regression model to describe the relationship between ADR (dependent variable) and predictor variables (hotel type, booking changes).

Data Preparation:

1. Extract the columns "hotel type," "number of booking changes," and "ADR" that are present in "hotel bookings" dataset.
2. To ensure that the dataset is clean and free of null or unnecessary entries for analysis, filter and clean it.

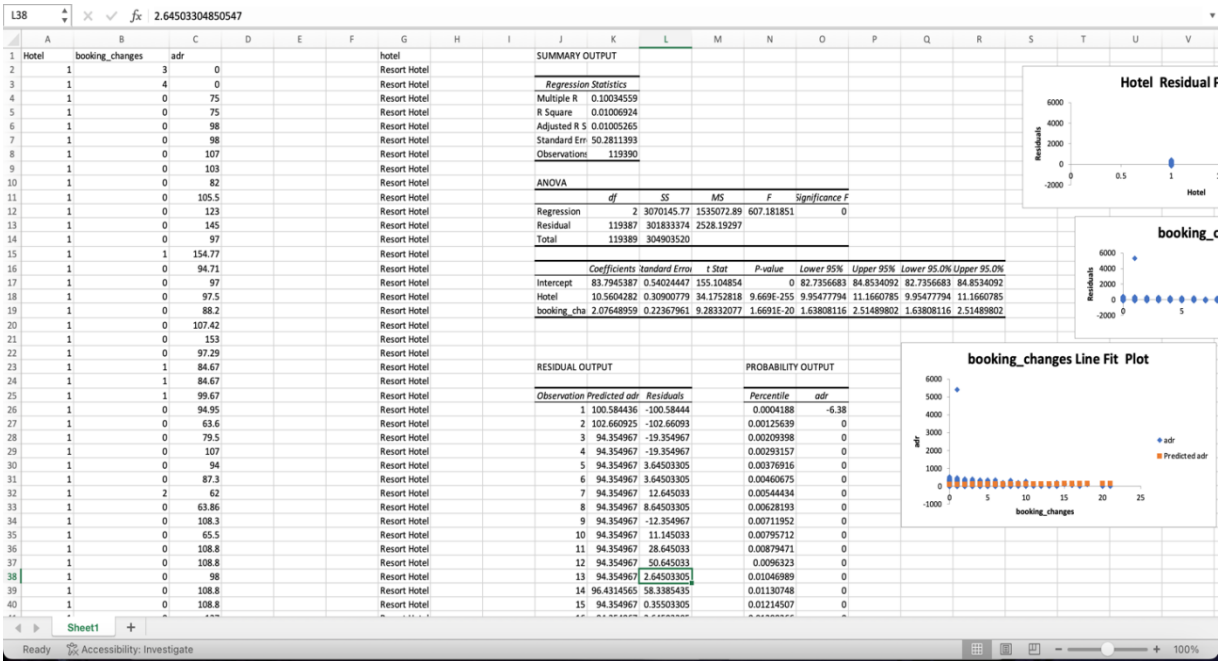
DATA ANALYTICS 1 (ADTA 5130) – FINAL PROJECT (GROUP 25)

Excel Data Arrangement:

1. Sort the data into columns, with the dependent variable being "ADR" and the independent variables being "hotel type," "number of booking changes".
2. A different observation or data point is represented by each row.

Conduct Linear Regression Analysis:

1. Open Excel and go to the 'Data' tab.Click on 'Data Analysis'.
2. Select 'Regression' from the list of available analysis tools.
3. As the dependent variable, enter the 'ADR' column, and the 'hotel type' and 'number of booking changes' columns as independent variables.
4. Other options (for example, confidence level) can be specified as needed.
5. Carry out the regression analysis.



LOGISTIC REGRESSION MODEL

Logistic Regression model to predict hotel reservation cancellations (binary outcome) based on predictors —lead time and days on the waiting list.

Data Preparation:

1. Assemble the "hotel_bookings" dataset by adding columns for predictors such as "lead time," "days on waiting list," and the binary result "is_canceled."
2. To ensure that the data is suitable for analysis, filter and clean the dataset to get rid of any null or unnecessary entries.

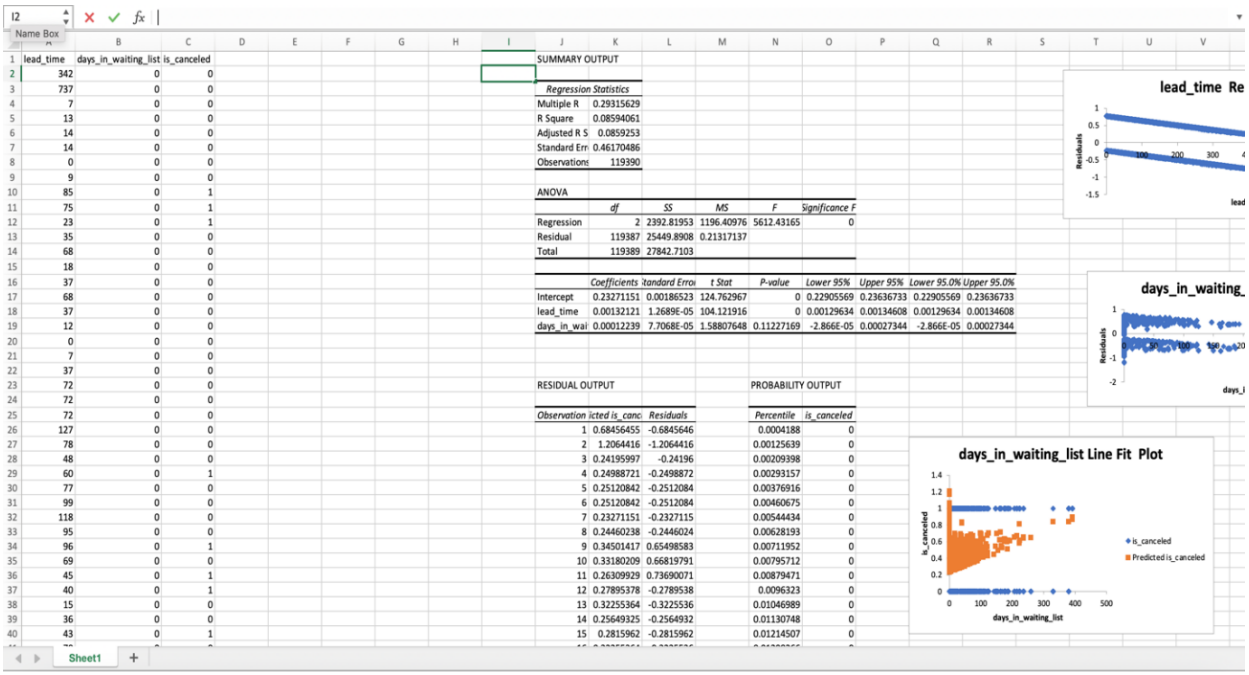
Prepare the Data in Excel:

1. Sort the data into columns, with the dependent variable being 'is_canceled' (cancellation status) and the independent variables being 'lead time', 'days on waiting list'.
2. A distinct observation or piece of data is represented by each row.

Logistic regression Analysis:

1. Launch Excel, then choose the 'Data' tab. Select 'Data Analysis'.
2. Select 'Regression' from the analytical tool list.Enter the lead time, days on waiting list, or other independent variables in the 'is_canceled' column as the dependent variable.
3. Indicate additional settings (like the confidence level) if required. Execute the analysis of linear regression.

DATA ANALYTICS 1 (ADTA 5130) – FINAL PROJECT (GROUP 25)



lead_time	days_in_waiting_list	is_canceled
342	0	0
737	0	0
7	0	0
13	0	0
14	0	0
14	0	0
0	0	0
9	0	0
85	0	1
75	0	1
23	0	1
35	0	0
68	0	0
18	0	0
37	0	0
68	0	0
37	0	0
12	0	0
0	0	0
7	0	0
37	0	0
72	0	0
72	0	0
72	0	0
127	0	0
78	0	0
48	0	0
60	0	1
77	0	0
99	0	0
118	0	0
95	0	0
96	0	1
69	0	0
45	0	1
40	0	1
15	0	0
36	0	0
43	0	1

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.29315629
R Square	0.08594061
Adjusted R S	0.0859253
Standard Err	0.46170486
Observations	119390

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2392.81953	1196.40976	5612.43165	0
Residual	119387	25449.8908	0.21317137		
Total	119389	27842.7103			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.23271151	0.00186523	124.762967	0	0.22905569	0.23636733	0.22905569	0.23636733
lead_time	0.00132121	1.2689E-05	104.121916	0	0.00129634	0.00134608	0.00129634	0.00134608
days_in_wai	0.00012239	7.7068E-05	1.58807648	0.11227169	-2.866E-05	0.00027344	-2.866E-05	0.00027344

RESIDUAL OUTPUT

Observation	*icted is_canc*	*Residuals*
1	0.68456455	-0.6845646
2	1.2064416	-1.2064416
3	0.24195997	-0.24196
4	0.24988721	-0.2498872
5	0.25120842	-0.2512084
6	0.25120842	-0.2512084
7	0.23271151	-0.2327115
8	0.24460238	-0.2446024
9	0.34501417	0.65498583
10	0.33180209	0.66819791
11	0.26309929	0.73690071
12	0.27895378	-0.2789538
13	0.32255364	-0.3225536
14	0.25649325	-0.2564932
15	0.2815962	-0.2815962

PROBABILITY OUTPUT

Percentile	*is_canceled*
0.0004188	0
0.00125639	0
0.00209398	0
0.00293157	0
0.00376916	0
0.00460675	0
0.00544434	0
0.00628193	0
0.00711952	0
0.00795712	0
0.00879471	0
0.0096323	0
0.01046989	0
0.01130748	0
0.01214507	0

POLYNOMIAL REGRESSION MODEL 3

Utilized Polynomial Regression to capture potential nonlinear relationships between the choice of agent and ADR.

Data Preparation:

1. Make sure the "hotel_bookings" dataset has columns for the Average Daily Rate (ADR) and the variable "agent," which represents the selections made by various agents.
2. To ensure that the dataset is clean and free of null or unnecessary entries for analysis, filter and clean it.

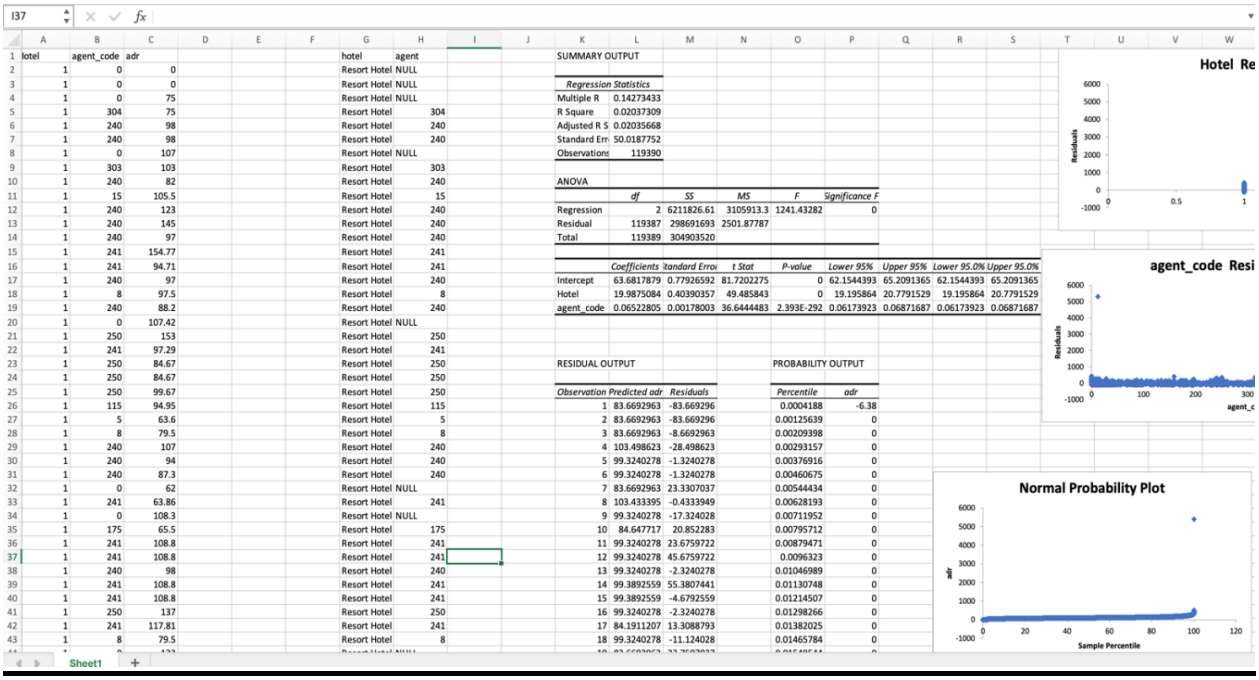
Excel Data Arrangement:

1. 'ADR' should be the dependent variable and 'agent' should be the independent variable in a columnar arrangement of the data.
2. A different observation or data point is represented by each row.

DATA ANALYTICS 1 (ADTA 5130) – FINAL PROJECT (GROUP 25)

Polynomial Regression Analysis:

1. Open Excel and choose the 'Data' tab.Select 'Data Analysis'.
2. Select 'Regression' from the toolkit for analysis.
3. Enter the agent as the independent variable and the ADR as the dependent variable.
4. Indicate the type of Regression.
5. Execute the analysis of regression.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.14273433
R Square	0.02037309
Adjusted R S	0.02035668
Standard Err	50.0187752
Observations	119390

ANOVA

	df	SS	MS	F	Significance F
Regression	2	6211826.61	3105913.3	1241.43282	0
Residual	119387	298691693	2501.87787		
Total	119389	304903520			

	Coefficients	tandard Erro	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	63.6817879	0.77926592	81.7202275	0	62.1544393	65.2091365	62.1544393	65.2091365
Hotel	19.9875084	0.40390357	49.485843	0	19.195864	20.7791529	19.195864	20.7791529
agent_code	0.06522805	0.00178003	36.6444483	2.393E-292	0.06173923	0.06871687	0.06173923	0.06871687

RESIDUAL OUTPUT

Observation	Predicted adr	Residuals
1	83.6692963	-83.669296
2	83.6692963	-83.669296
3	83.6692963	-8.6692963
4	103.498623	-28.498623
5	99.3240278	-1.3240278
6	99.3240278	-1.3240278
7	83.6692963	23.3307037
8	103.433395	-0.4333949
9	99.3240278	-17.324028
10	84.647717	20.852283
11	99.3240278	23.6759722
12	99.3240278	45.6759722
13	99.3240278	-2.3240278
14	99.3892559	55.3807441
15	99.3892559	-4.6792559
16	99.3240278	-2.3240278
17	84.1911207	13.3088793
18	99.3240278	-11.124028

PROBABILITY OUTPUT

Percentile	adr
0.0004188	-6.38
0.00125639	0
0.00209398	0
0.00293157	0
0.00376916	0
0.00460675	0
0.00544434	0
0.00628193	0
0.00711952	0
0.00795712	0
0.00879471	0
0.0096323	0
0.01046989	0
0.01130748	0
0.01214507	0
0.01298266	0
0.01382025	0
0.01465784	0

RESULTS

ONE-WAY ANOVA MODEL

J	K	L	M	N	O	P
Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
arrival_month_code	119390	782301	6.55248346	9.55192387		
adr	119390	12157617.6	101.831122	2553.8661		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	541912337	1	541912337	422804.499	0	3.84149777
Within Groups	306043914	238778	1281.70901			
Total	847956251	238779				

Notable Differences Across Arrival Months:

1. The ANOVA test reveals a very noteworthy F-statistic (F = 422804.499) for the "arrival_month_code" in relation to ADR (Average Daily Rate).
2. Significant variation in ADR across different arrival months is suggested by the between-groups variance (SS), which is noticeably larger than the within-groups variance.

Statistical significance is indicated by a low P-value:

1. The ANOVA yielded a p-value of nearly zero ($p < 0.001$), indicating compelling evidence opposing the null hypothesis.
2. shows that it is extremely unlikely that variations in ADR between arrival months happened by accident.

Significant Disparities in ADR:

1. The incredibly low p-value rules out the null hypothesis, which states that there is no variation in ADR between arrival months.
2. Implies that there may be notable variations in ADR between at least some of the assessed months.

F-Statistics marked differences between groups:

1. More proof that the null hypothesis is false is provided by the F-statistic value being significantly greater than the critical value (F crit).
2. Verifies that the ADR varies significantly depending on the arrival month.

These findings, which show highly significant differences in ADR between arrival months and support the rejection of the null hypothesis, highlight the importance of the ANOVA results. They also indicate significant differences in ADR between the months that were examined.

TWO-WAY ANOVA MODEL

	2	2	7	3.5	4.5		
	2	2	7	3.5	4.5		
	2	2	9	4.5	12.5		
	stays_in_week_nights	119390	298511	2.50030153	3.64155399		
	stays_in_weekend_nigh	119390	110746	0.92759863	0.99722891		
	ANOVA						
	Source of Variation	SS	df	MS	F	P-value	F crit
	Rows	390431.432	119389	3.27024627	2.38959353	0	1.00956634
	Columns	147649.281	1	147649.281	107888.439	0	3.84153671
	Error	163388.219	119389	1.36853663			
	Total	701468.932	238779				

This table shows the results of a two-factor ANOVA where the factors are 'stays_in_week_nights' and 'stays_in_weekend_nights'. Here's the breakdown:

1. **Rows (stays_in_week_nights):**
   - **SS (Sum of Squares):** 390431.43
   - **df (Degrees of Freedom):** 119389
   - **MS (Mean Square):** 3.2702
   - **F-value:** 2.3896
   - **P-value:** 0.0 (approx.)
   - **Interpretation:** The factor 'stays_in_week_nights' shows a statistically significant effect on the dependent variable ($p < 0.05$).
2. **Columns (stays_in_weekend_nights):**
   - **SS (Sum of Squares):** 147649.28
   - **df (Degrees of Freedom):** 1
   - **MS (Mean Square):** 147649.28

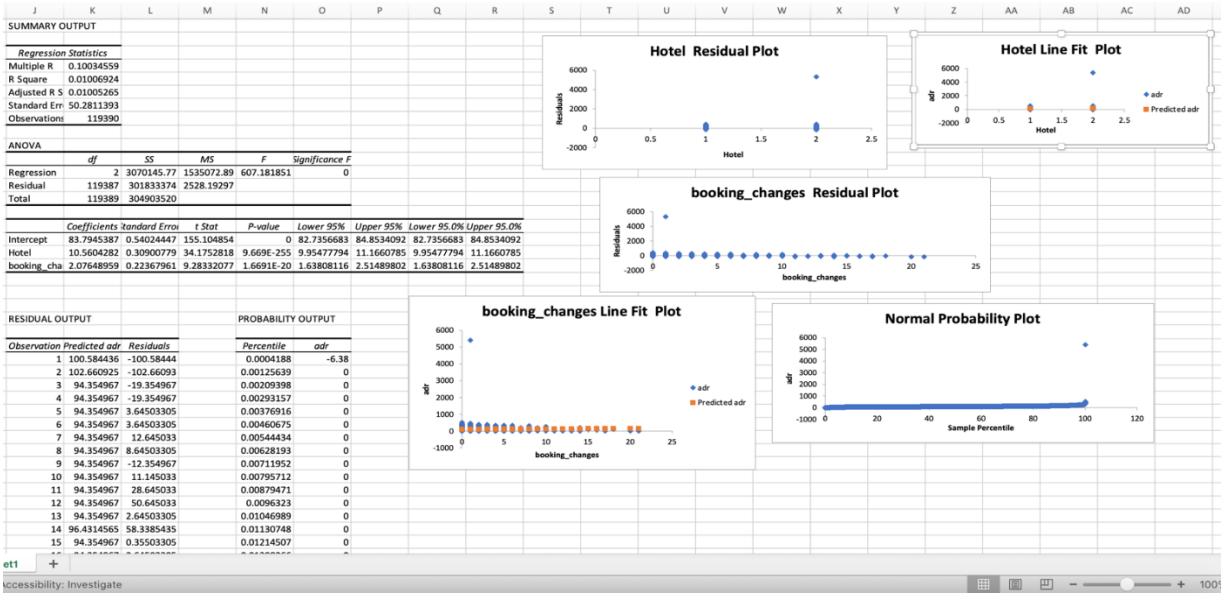
- **F-value:** 107888.44
- **P-value:** 0.0 (approx.)
- **Interpretation:** The factor 'stays_in_weekend_nights' demonstrates a highly significant effect on the dependent variable.

3. **Error:**
   - **SS (Sum of Squares):** 163388.22
   - **df (Degrees of Freedom):** 119389
   - **MS (Mean Square):** 1.3685
4. **Total:**
   - **SS (Sum of Squares):** 701468.93
   - **df (Degrees of Freedom):** 238779

Interpretation:

1. With extremely low p-values ($p < 0.05$), both factors, "stays_in_weeknights" and "stays_in_weekend_nights," show significant effects on the dependent variable.
2. This table does not specifically display the interaction effect between these factors.
3. Based on the p-values given in the analysis, this ANOVA output indicates that "stays_in_weeknights" and "stays_in_weekend_nights" have a statistically significant impact on the dependent variable.

LINEAR REGRESSION MODEL



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.10034559
R Square	0.01006924
Adjusted R S	0.01005265
Standard Err	50.2811393
Observations	119390

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	3070145.77	1535072.89	607.181851	0
Residual	119387	301833374	2528.19297		
Total	119389	304903520			

	Coefficients	*tandard Erroi*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	83.7945387	0.54024447	155.104854	0	82.7356683	84.8534092	82.7356683	84.8534092
Hotel	10.5604282	0.30900779	34.1752818	9.669E-255	9.95477794	11.1660785	9.95477794	11.1660785
booking_cha	2.07648959	0.22367961	9.28332077	1.6691E-20	1.63808116	2.51489802	1.63808116	2.51489802

RESIDUAL OUTPUT

Observation	*Predicted adr*	*Residuals*
1	100.584436	-100.58444
2	102.660925	-102.66093
3	94.354967	-19.354967
4	94.354967	-19.354967
5	94.354967	3.64503305
6	94.354967	3.64503305
7	94.354967	12.645033
8	94.354967	8.64503305
9	94.354967	-12.354967
10	94.354967	11.145033
11	94.354967	28.645033
12	94.354967	50.645033
13	94.354967	2.64503305
14	96.4314565	58.3385435
15	94.354967	0.35503305

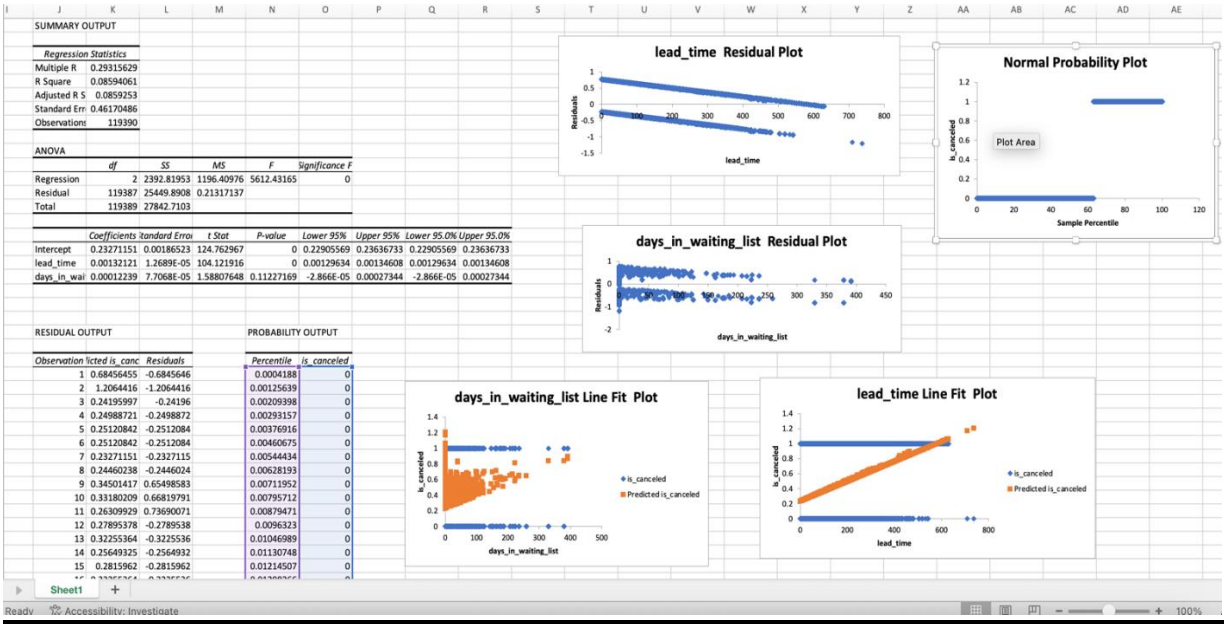
PROBABILITY OUTPUT

Percentile	*adr*
0.0004188	-6.38
0.00125639	0
0.00209398	0
0.00293157	0
0.00376916	0
0.00460675	0
0.00544434	0
0.00628193	0
0.00711952	0
0.00795712	0
0.00879471	0
0.0096323	0
0.01046989	0
0.01130748	0
0.01214507	0

The ADR (Average Daily Rate) and predictors like hotel type and booking changes are compared

using a regression model. The model's low p-values suggest that changes in booking and hotel type have a statistically significant effect on ADR. Even though these predictors are significant, their ability to explain variations in ADR is limited, as the entire model only accounts for roughly 1% of the variation in ADR.

LOGISTIC REGRESSION MODEL



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.29315629
R Square	0.08594061
Adjusted R S	0.0859253
Standard Err	0.46170486
Observations	119390

ANOVA

	df	SS	MS	F	Significance F
Regression	2	2392.81953	1196.40976	5612.43165	0
Residual	119387	25449.8908	0.21317137		
Total	119389	27842.7103			

	Coefficients	Standard Erroi	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.23271151	0.00186523	124.762967	0	0.22905569	0.23636733	0.22905569	0.23636733
lead_time	0.00132121	1.2689E-05	104.121916	0	0.00129634	0.00134608	0.00129634	0.00134608
days_in_wai	0.00012239	7.7068E-05	1.58807648	0.11227169	-2.866E-05	0.00027344	-2.866E-05	0.00027344

RESIDUAL OUTPUT

Observation	Predicted is_canc	Residuals
1	0.68456455	-0.6845646
2	1.2064416	-1.2064416
3	0.24195997	-0.24196
4	0.24988721	-0.2498872
5	0.25120842	-0.2512084
6	0.25120842	-0.2512084
7	0.23271151	-0.2327115
8	0.24460238	-0.2446024
9	0.34501417	0.65498583
10	0.33180209	0.66819791
11	0.26309929	0.73690071
12	0.27895378	-0.2789538
13	0.32255364	-0.3225536
14	0.25649325	-0.2564932
15	0.2815962	-0.2815962

PROBABILITY OUTPUT

Percentile	is_canceled
0.0004188	0
0.00125639	0
0.00209398	0
0.00293157	0
0.00376916	0
0.00460675	0
0.00544434	0
0.00628193	0
0.00711952	0
0.00795712	0
0.00879471	0
0.0096323	0
0.01046989	0
0.01130748	0
0.01214507	0

Model Fit and Explanation:

Using lead time and days_in_waiting_list as predictors, the regression model accounts for about 8.59% of the variability in predicting "is_canceled".

Important Predictors:

1. Lead_time has a substantial influence on predicting cancellation probability, as evidenced by its highly significant relationship ($p < 0.001$) with "is_canceled."
2. Lead time appears to have a greater influence on cancellation prediction than days_in_waiting_list, as there is no statistically significant relationship between the two ($p = 0.112$).

The impact of the predictors:

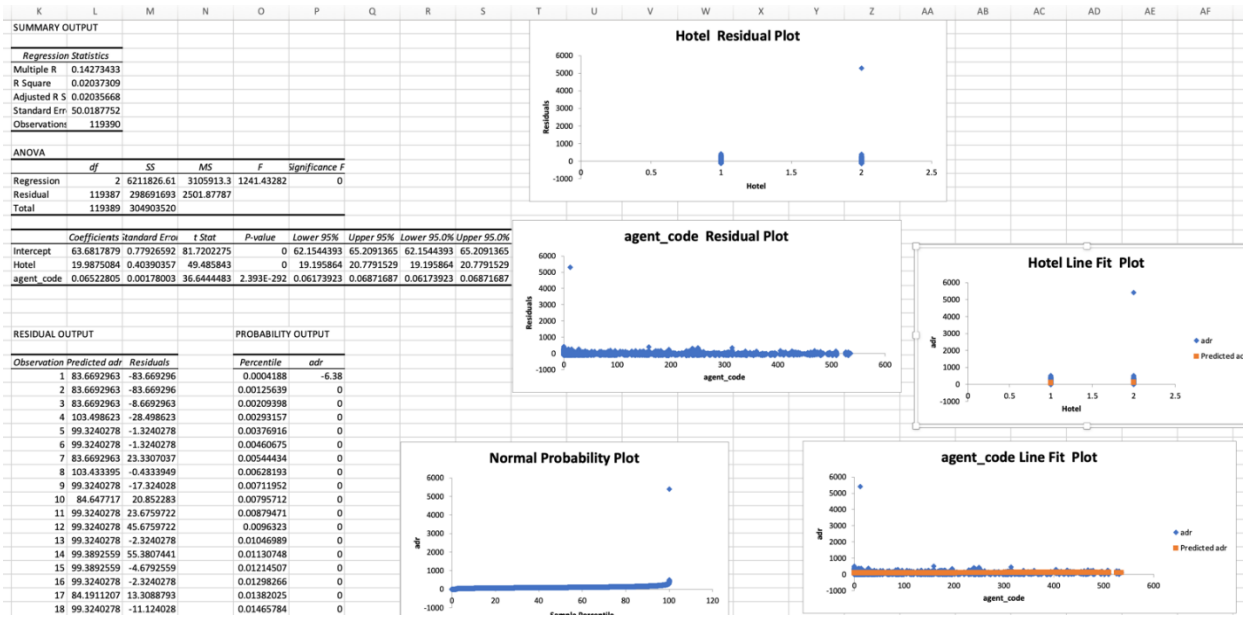
1. The probability of cancellation increases by about 0.00132 for every unit increase in lead_time.
2. Days_in_waiting_list may not have much of an effect on cancellation probabilities if its p-

value is above the significance threshold.

Reliability Analysis:

1. The residuals show differences between the observed and predicted "is_canceled" values, pointing to possible shortcomings in the model's ability to reliably predict cancellations.
2. The main conclusions drawn from the regression analysis are summarized in these points, which highlight the model's ability to explain "is_canceled," the importance of predictors, and possible limits on prediction accuracy.

POLYNOMIAL REGRESSION MODEL



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.14273433
R Square	0.02037309
Adjusted R S	0.02035668
Standard Erro	50.0187752
Observations	119390

ANOVA

	df	SS	MS	F	Significance F
Regression	2	6211826.61	3105913.3	1241.43282	0
Residual	119387	298691693	2501.87787		
Total	119389	304903520			

	Coefficients	Standard Erroi	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	63.6817879	0.77926592	81.7202275	0	62.1544393	65.2091365	62.1544393	65.2091365
Hotel	19.9875084	0.40390357	49.485843	0	19.195864	20.7791529	19.195864	20.7791529
agent_code	0.06522805	0.00178003	36.6444483	2.393E-292	0.06173923	0.06871687	0.06173923	0.06871687

RESIDUAL OUTPUT

Observation	Predicted adr	Residuals
1	83.6692963	-83.669296
2	83.6692963	-83.669296
3	83.6692963	-8.6692963
4	103.498623	-28.498623
5	99.3240278	-1.3240278
6	99.3240278	-1.3240278
7	83.6692963	23.3307037
8	103.433395	-0.4333949
9	99.3240278	-17.324028
10	84.647717	20.852283
11	99.3240278	23.6759722
12	99.3240278	45.6759722
13	99.3240278	-2.3240278
14	99.3892559	55.3807441
15	99.3892559	-4.6792559
16	99.3240278	-2.3240278
17	84.1911207	13.3088793
18	99.3240278	-11.124028

PROBABILITY OUTPUT

Percentile	adr
0.0004188	-6.38
0.00125639	0
0.00209398	0
0.00293157	0
0.00376916	0
0.00460675	0
0.00544434	0
0.00628193	0
0.00711952	0
0.00795712	0
0.00879471	0
0.0096323	0
0.01046989	0
0.01130748	0
0.01214507	0
0.01298266	0
0.01382025	0
0.01465784	0

Model Fit and Explanation:

Using predictors such as Hotel type and Agent code, the regression model accounts for about 2% of the variability in Average Daily Rate (ADR).

Important Predictors:

Hotel type and Agent code show a highly significant statistical relationship ($p < 0.001$) with ADR, suggesting that they have a substantial influence on ADR values.

Effect of Predictors:

An average ADR increase of 19.99 is linked to a unit change in hotel type, whereas an average ADR increase of 0.065 is associated with each additional unit in agent code.

Analysis of the Residuals:

The existence of extreme negative residuals (-83.67) indicates possible outliers in which the model's predictions substantially differ from the observed ADR values, indicating the need for additional research.

CONCLUSION:

The study examined factors influencing Average Daily Rates (ADR) for hotel reservations made between 2015 and 2017. Even though they demonstrated statistical significance in forecasting ADR, variables like hotel type and booking modifications could only account for a small portion of its variability. This implies that ADR may be influenced by additional unaccounted-for variables. To understand the larger dynamics underlying ADR fluctuations in hotel bookings, more research is necessary.