Assignment 3 - Strategic Marketing Decision Making 2025-2026

Gonçalo Passinhas (732518)

2025-10-07

Question 1:

Although the Elbow plot does not show a sharply defined bend, it reveals that the reduction in the total within-cluster sum of squares begins to level off around k = 4, suggesting that additional clusters beyond this point yield only limited improvements in model fit. This gradual flattening of the curve indicates that four clusters provide a good trade-off between minimizing within-cluster variance and maintaining a parsimonious solution. The Silhouette plot further supports this finding, as the average silhouette width reaches its maximum at k = 4, indicating that this solution achieves the strongest balance between cluster cohesion (similarity within clusters) and separation (differences between clusters). Considering the evidence from both plots, a four-cluster solution appears to be the most appropriate choice for segmenting hotel customers based on the selected variables.

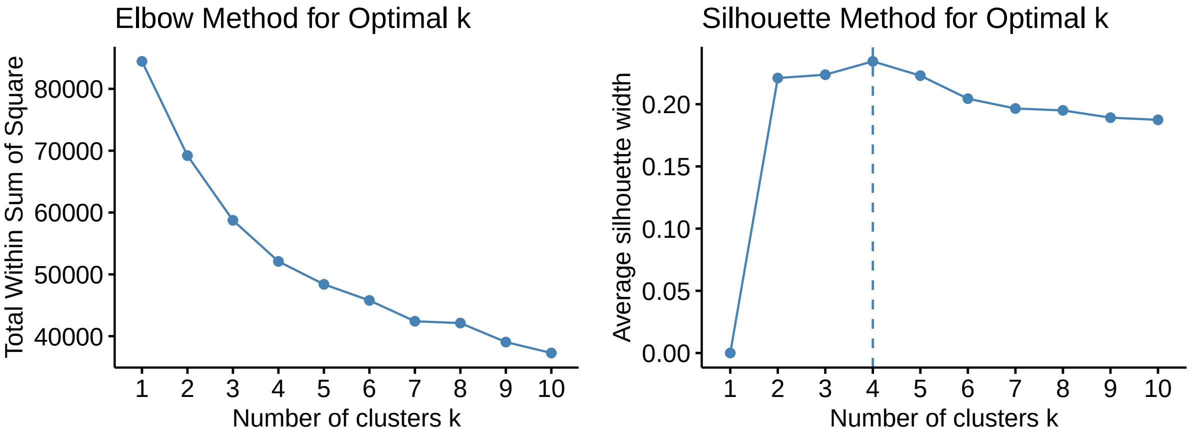

Figure 1: Elbow (left) and Silhouette (right) Method for Optimal k

Question 2:

Based on the k-means cluster analysis with four clusters, two bar plots were generated to illustrate the segmentation results. As shown in Figure 2, the cluster sizes vary substantially, with Cluster 3 (3,504 customers) and Cluster 4 (3,379 customers) representing the largest segments, while Cluster 1 (998 customers) and Cluster 2 (1,503 customers) are considerably smaller. This uneven distribution suggests that certain customer profiles are more common within the dataset.

Figure 3 presents the overall segmentation solution by displaying the mean values of the main variables for each cluster. The figure highlights clear differences across clusters in terms of booking behaviour and revenue indicators. For example, some clusters exhibit higher lodging and other revenue values, while others are characterized by shorter lead times or fewer room nights. These variations indicate distinct customer

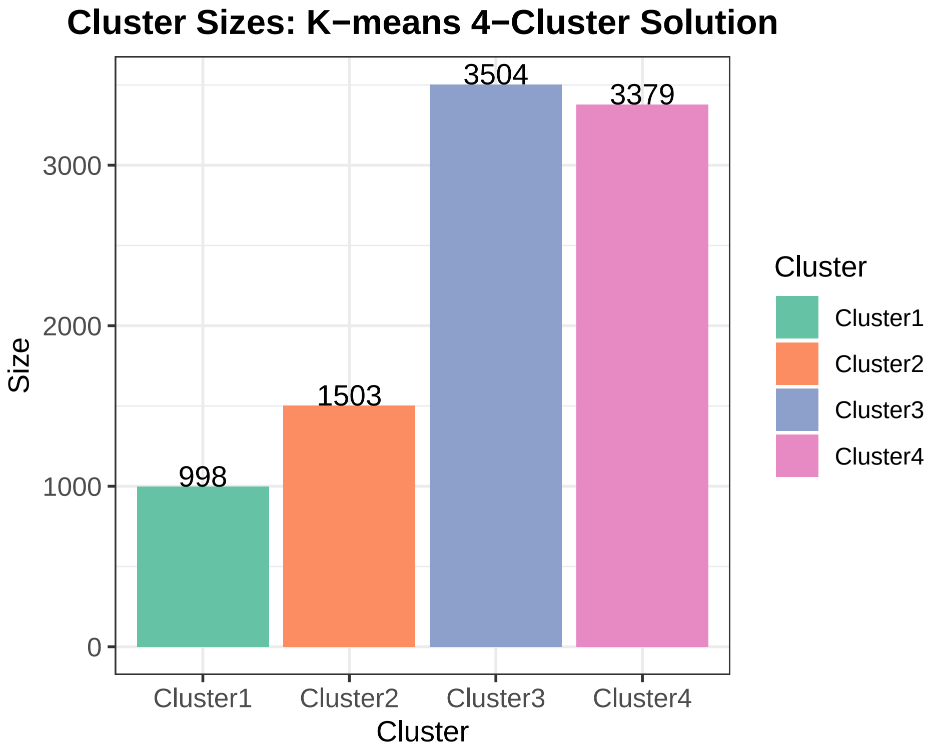

Figure 2: Cluster Sizes for the K-means 4-Cluster Solution

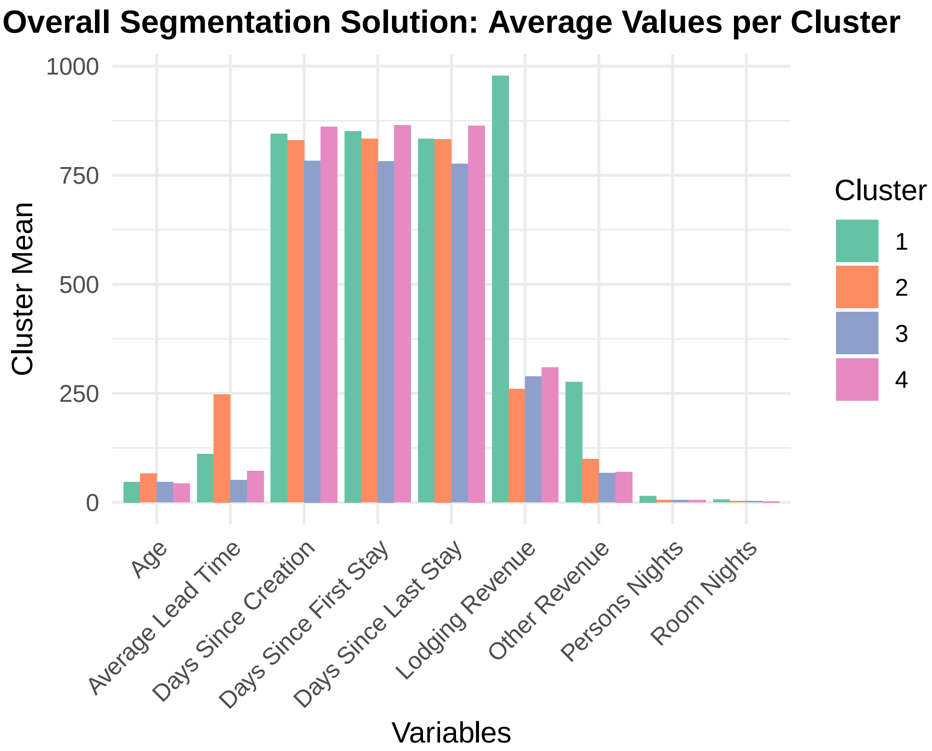

Figure 3: Overall Segmentation Solution: Mean Values of Key Variables Across the Four Clusters

groups that differ in their spending patterns, booking habits, and engagement with the hotel. Overall, the four-cluster solution provides a meaningful segmentation that can be used to tailor marketing and retention strategies to specific customer segments.

Question 3:

Management is particularly interested in understanding which segments drive lodging revenue and which contribute most to ancillary (other) revenues. To answer this, we computed the mean lodging revenue and mean other revenue for each cluster and visualised the results with bar charts.

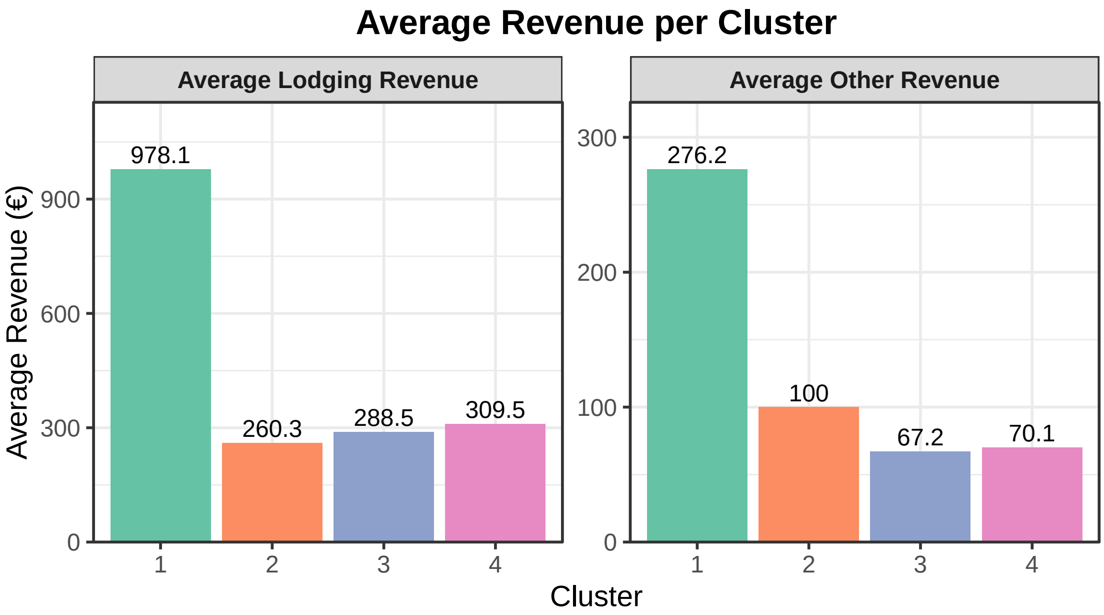

Figure 4: Average Lodging Revenue per Cluster (left) and Average Other Revenue per Cluster (right)

The bar charts clearly show that **Cluster 1** generates the highest lodging revenue (≈ €980 on average) and also contributes the most to ancillary revenue (≈ €276). These customers tend to be middle-aged and stay for many nights (nearly 15 person-nights on average). In comparison, Clusters 3 and 4 exhibit lodging revenues around €290–€310 and ancillary revenues of €67–€70, while Cluster 2 (with the highest average lead time and the oldest customers) has even lower lodging and other revenue values. The substantial gap between Cluster 1 and the other groups indicates that targeting this segment can have a disproportionate impact on overall hotel revenue. The definitions of **LodgingRevenue** and **OtherRevenue** —total spending on room-related expenses and total spending on food, beverage, spa and other services —underline the managerial relevance of these metrics.

Question 4:

Management also requested guidance on how to position the hotel's services for the selected target segment(s). Building on the cluster profiles and revenue analysis, two actionable recommendations are proposed.

1. **Cultivate and reward the high-value segment (Cluster 1).** Customers in Cluster 1 are valuable: they spend almost four times more on lodging and ancillary services than the average guest, tend to be in their forties, and book roughly 110 days in advance. They also stay much longer (≈ 15 person-nights)

than other segments. To retain and further monetise this group, the hotel should introduce a **premium loyalty program** that offers personalised perks (e.g., room upgrades, dedicated concierge support and priority check-in) and pre-arrival upselling communications. Because their long lead times provide an extended window before arrival, the hotel can send targeted emails promoting spa packages, dining experiences and tour add-ons. Bundling these services as "all-inclusive" or "experience" packages encourages guests to commit before arrival, increasing ancillary revenue while enhancing perceived value.

2. **Develop tailored packages for the senior planner segment (Cluster 2).** Cluster 2 is characterised by the oldest customers (around 66 years on average) and the longest booking horizon ($\approx$ 248 days), yet they generate relatively modest lodging and ancillary revenue. Their early commitment provides an opportunity to influence purchase behaviour. The hotel should design **senior-friendly, themed packages**—for example, wellness retreats, cultural tours or off-peak "quiet-season" stays—with transparent pricing and flexible cancellation policies. To increase ancillary spending, these packages should incorporate dining vouchers, guided activities and spa treatments. Because older travellers often value personal contact, follow-up calls or personalised emails after booking can be used to cross-sell upgrades or services aligned with their interests and mobility needs. These efforts can elevate the revenue contribution of this sizable segment without significantly increasing acquisition costs.

Taken together, these recommendations leverage the cluster analysis to focus resources on segments that drive revenue today while nurturing segments with potential for growth. Targeting high-value guests with personalised incentives safeguards existing revenue streams, whereas curated packages for senior planners tap into a large, currently under-monetised group. Both initiatives align with the hotel's objective of increasing lodging and ancillary revenues and demonstrate how data-driven segmentation can inform strategic marketing decisions.

Appendix

This appendix contains:

1. The complete R code used in this assignment for reproducibility.
2. Supporting tables for Question 4 (cluster statistics and revenue shares).

```r
# --- Import data ---
hotel <- read.csv("HotelCustomersSubset_sample.csv")

# --- Load packages ---
library(tidyverse)
library(cluster)
library(factoextra)
library(tidyr)
library(ggplot2)
library(dplyr)
library(gt)
# --- Set seed for reproducibility ---
set.seed(123)

# First, ensure the Age column is numeric
# If there are text values like "NULL", they will become NA
hotel$Age <- as.numeric(as.character(hotel$Age))
hotel <- hotel %>%
  mutate(across(everything(), as.numeric))


# --- Data cleaning ---
hotel <- hotel %>%
  filter(!is.na(Age),
         Age >= 18 & Age <= 95,
         Age != "NULL")

# --- Select relevant variables ---
hotel <- hotel %>%
  select(Age, AverageLeadTime, DaysSinceCreation,
         LodgingRevenue, OtherRevenue, PersonsNights,
         RoomNights, DaysSinceLastStay, DaysSinceFirstStay)

# --- Scale the data ---
hotel_scaled <- scale(hotel)
# --- Figure 1: Elbow Method ---
fviz_nbclust(hotel_scaled, kmeans, method = "wss") +
  ggtitle("Elbow Method for Optimal k")

# --- Figure 2: Silhouette Method ---
fviz_nbclust(hotel_scaled, kmeans, method = "silhouette") +
  ggtitle("Silhouette Method for Optimal k")
## --- Question 2: K-means Cluster Analysis with k = 4 ---

set.seed(123)
```

```r
### Let's assume 4 clusters (based on Question 1 results)

HotelCluster_4k <- kmeans(hotel_scaled, 4)

# Check the number of observations in each cluster
HotelCluster_4k[["size"]]

sizes4k <- data.frame(Size = HotelCluster_4k[["size"]],
                      Cluster = c("Cluster1", "Cluster2", "Cluster3", "Cluster4"))

## Let's get the means of each variable in each cluster
hotel$k4Cluster <- HotelCluster_4k[["cluster"]]

library(dplyr)

summarystats.percluster_4k <- hotel %>%
  group_by(k4Cluster) %>%
  summarise_if(is.numeric, mean, na.rm = TRUE)

head(summarystats.percluster_4k)

## Let's plot the clusters on a two-dimensional space, using PCA
library(factoextra)

fviz_cluster(HotelCluster_4k, data = hotel_scaled, ellipse.type = "norm") +
  ggtitle("Customer Segmentation: 4-Cluster Solution (PCA Projection)")

### Let's plot the 4-cluster solution
library(ggplot2)

# Define colors similar to the segmentation plot (Set2 palette)
cluster_colors <- c("#66C2A5", "#FC8D62", "#8DA0CB", "#E78AC3")

ggplot(sizes4k, aes(x = factor(Cluster), y = Size, fill = factor(Cluster))) +
  geom_col() +
  geom_text(aes(label = Size), vjust = 0) +
  xlab("Cluster") +
  ylab("Size") +
  ggtitle("Cluster Sizes: K-means 4-Cluster Solution") +
  scale_fill_manual(values = cluster_colors, name = "Cluster") +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 11)
  )
# --- Calculate means per cluster, excluding the cluster label itself ---
summarystats.percluster_4k <- hotel %>%
  group_by(k4Cluster) %>%
  summarise(across(where(is.numeric) & !matches("k4Cluster"), mean, na.rm = TRUE))

# --- Rename variables for cleaner display ---
summarystats.percluster_4k <- summarystats.percluster_4k %>%
  rename(
    "Age" = Age,
    "Average Lead Time" = AverageLeadTime,
    "Days Since Creation" = DaysSinceCreation,
    "Lodging Revenue" = LodgingRevenue,
    "Other Revenue" = OtherRevenue,
    "Persons Nights" = PersonsNights,
    "Room Nights" = RoomNights,
    "Days Since Last Stay" = DaysSinceLastStay,
    "Days Since First Stay" = DaysSinceFirstStay
  )

# --- Reshape data for plotting ---
summarystats_long_4k <- summarystats.percluster_4k %>%
  pivot_longer(-k4Cluster, names_to = "Variable", values_to = "Mean")

# --- Bar plot: Overall segmentation solution ---
ggplot(summarystats_long_4k, aes(x = Variable, y = Mean, fill = factor(k4Cluster))) +
  geom_col(position = "dodge") +
  theme_minimal() +
  xlab("Variables") +
  ylab("Cluster Mean") +
  ggtitle("Overall Segmentation Solution: Average Values per Cluster") +
  scale_fill_brewer(palette = "Set2", name = "Cluster") +
   theme(
    plot.title = element_text(size=12, hjust = 0.5, face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.title = element_text(size = 11)
  )

# Compute revenue statistics per cluster
revenue_stats <- hotel %>%
  group_by(k4Cluster) %>%
  summarise(
    MeanLodgingRevenue = mean(LodgingRevenue, na.rm = TRUE),
    MeanOtherRevenue   = mean(OtherRevenue,   na.rm = TRUE),
    .groups = "drop"
  ) %>%
  mutate(Cluster = factor(k4Cluster))

# Long format for faceting
revenue_long <- revenue_stats %>%
  select(Cluster, MeanLodgingRevenue, MeanOtherRevenue) %>%
  pivot_longer(
    cols = c(MeanLodgingRevenue, MeanOtherRevenue),
    names_to = "Metric",
    values_to = "Mean"
  ) %>%
  mutate(Metric = recode(Metric,
                         "MeanLodgingRevenue" = "Average Lodging Revenue",
                         "MeanOtherRevenue"   = "Average Other Revenue"))

# One clean, readable plot
```

```r
ggplot(revenue_long, aes(x = Cluster, y = Mean, fill = Cluster)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label = round(Mean, 1)), vjust = -0.35, size = 3) +
  facet_wrap(~ Metric, ncol = 2, scales = "free_y") +
  labs(
    title = "Average Revenue per Cluster",
    x = "Cluster",
    y = "Average Revenue (\u20AC)"
  ) +
  scale_fill_manual(values = c("#66C2A5", "#FC8D62", "#8DA0CB", "#E78AC3")) +
  # Headroom for labels + no clipping
  scale_y_continuous(expand = expansion(mult = c(0, 0.18))) +
  coord_cartesian(clip = "off") +
  theme_bw(base_size = 11) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    strip.text = element_text(face = "bold"),
    plot.margin = margin(10, 12, 16, 10) # top, right, bottom, left
  )
# Helper for 95% CI of the mean
mean_ci <- function(x) {
  x <- x[is.finite(x)]
  n <- length(x)
  m <- mean(x); s <- sd(x); se <- s / sqrt(n)
  me <- qt(0.975, df = max(n - 1, 1)) * se
  c(mean = m, lwr = m - me, upr = m + me, n = n)
}

vars_of_interest <- c(
  "Age","AverageLeadTime","PersonsNights","RoomNights",
  "LodgingRevenue","OtherRevenue","DaysSinceCreation",
  "DaysSinceLastStay","DaysSinceFirstStay"
)

cluster_summary_long <- hotel %>%
  mutate(k4Cluster = factor(k4Cluster)) %>%
  pivot_longer(all_of(vars_of_interest), names_to = "Variable", values_to = "Value") %>%
  group_by(k4Cluster, Variable) %>%
  summarise(
    Mean = mean(Value, na.rm = TRUE),
    Median = median(Value, na.rm = TRUE),
    SD = sd(Value, na.rm = TRUE),
    N = sum(is.finite(Value)),
    CI_Lower = mean_ci(Value)["lwr"],
    CI_Upper = mean_ci(Value)["upr"],
    .groups = "drop"
  )

cluster_summary_table <- cluster_summary_long %>%
  mutate(
    Mean = round(Mean, 2),
    Median = round(Median, 2),
    SD = round(SD, 2),
    CI = paste0("[", round(CI_Lower, 2), ", ", round(CI_Upper, 2), "]")
  ) %>%
  select(k4Cluster, Variable, N, Mean, Median, SD, CI) %>%
  arrange(Variable, k4Cluster)

# Build the table
cluster_table <- cluster_summary_table %>%
  gt(rowname_col = "Variable", groupname_col = "k4Cluster") %>%
  tab_header(title = "Appendix - Cluster Profiles (Means, Medians, 95% CIs)")

invisible(NULL)  # ensure nothing prints here
cluster_table
```

Appendix –Cluster Profiles (Means, Medians, 95% CIs)

	N	Mean	Median	SD	CI
1					
Age	998	47.11	49.00	13.76	[46.25, 47.96]
AverageLeadTime	998	110.62	115.00	69.72	[106.29, 114.95]
DaysSinceCreation	998	845.21	853.00	36.90	[842.92, 847.51]
DaysSinceFirstStay	998	851.29	859.00	37.87	[848.94, 853.64]
DaysSinceLastStay	998	833.18	857.00	110.94	[826.28, 840.07]
LodgingRevenue	998	978.10	790.20	591.97	[941.33, 1014.88]
OtherRevenue	998	276.24	206.50	272.11	[259.33, 293.14]
PersonsNights	998	14.98	14.00	7.04	[14.54, 15.41]
RoomNights	998	6.52	6.00	5.06	[6.2, 6.83]
2					
Age	1503	65.90	67.00	11.55	[65.31, 66.48]
AverageLeadTime	1503	247.85	245.00	83.51	[243.62, 252.07]
DaysSinceCreation	1503	830.33	827.00	30.12	[828.8, 831.85]
DaysSinceFirstStay	1503	833.21	830.00	29.89	[831.7, 834.73]
DaysSinceLastStay	1503	832.80	830.00	31.64	[831.19, 834.4]
LodgingRevenue	1503	260.34	225.00	172.82	[251.59, 269.08]
OtherRevenue	1503	99.97	90.00	69.45	[96.45, 103.48]
PersonsNights	1503	5.47	4.00	2.66	[5.33, 5.6]
RoomNights	1503	2.92	3.00	1.16	[2.86, 2.98]
3					
Age	3504	46.59	46.00	12.81	[46.17, 47.02]
AverageLeadTime	3504	51.42	35.00	55.41	[49.58, 53.26]
DaysSinceCreation	3504	783.47	782.00	22.65	[782.72, 784.22]
DaysSinceFirstStay	3504	782.18	785.00	57.30	[780.28, 784.07]
DaysSinceLastStay	3504	776.35	784.00	76.22	[773.83, 778.88]
LodgingRevenue	3504	288.54	254.00	187.39	[282.33, 294.75]
OtherRevenue	3504	67.18	48.00	64.00	[65.06, 69.3]
PersonsNights	3504	5.23	6.00	3.07	[5.13, 5.33]
RoomNights	3504	2.78	3.00	1.37	[2.74, 2.83]
4					
Age	3379	43.19	44.00	12.38	[42.77, 43.61]
AverageLeadTime	3379	71.51	51.00	64.94	[69.32, 73.7]
DaysSinceCreation	3379	861.41	861.00	22.70	[860.65, 862.18]
DaysSinceFirstStay	3379	863.96	863.00	22.79	[863.2, 864.73]
DaysSinceLastStay	3379	863.71	863.00	23.57	[862.91, 864.5]
LodgingRevenue	3379	309.54	283.05	178.52	[303.52, 315.56]
OtherRevenue	3379	70.07	56.00	61.23	[68, 72.13]
PersonsNights	3379	5.49	6.00	3.03	[5.39, 5.59]
RoomNights	3379	2.56	2.00	1.28	[2.52, 2.6]