**פרויקט בקורס רגרסיה ליניארית**

**חלק ב'**



תוכן עניינים

[1. תקציר מנהלים: 3](#_Toc107322994)

[2. עיבוד מקדים: 4](#_Toc107322995)

[2.1 הסרה של משתנים: 4](#_Toc107322996)

[2.2 התאמת משתנים: 8](#_Toc107322997)

[2.3 הגדרת משתנה דמה: 9](#_Toc107322998)

[2.4 משתני אינטראקציה: 10](#_Toc107322999)

[3. התאמת המודל ובדיקת הנחות המודל: 11](#_Toc107323000)

[3.1 בחירת משתני המודל: 11](#_Toc107323001)

[3.2 בדיקת הנחות המודל: 11](#_Toc107323002)

[4. שיפור המודל: 14](#_Toc107323003)

[נספחים 15](#_Toc107323004)

[אינטראקציות נוספות שנבדקו : 15](#_Toc107323005)

[בניית המודל : 17](#_Toc107323006)

[בדיקת הנחות מודלים לשיפור : 23](#_Toc107323007)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **סוג המשתנה** | **סימון במודל** | **יחידת מידה** | **סוג המשתנה** | **הסבר קצר על המשתנה** |
| **מוסבר Y** | adr | דולר | רציף | תעריף יומי ממוצע -  מחושב על ידי חלוקת סכום כל עסקאות הלינה במספר הכולל של לילות השהייה |
| **מסביר X1** | lead\_time | מספר שלם | רציף | מספר הימים שחלפו בין תאריך ההזמנה לבין תאריך ההגעה. |
| **מסביר X2** | arrival\_date\_month | טקסטואלי | קטגוריאלי | חודש ההגעה |
| **מסביר X3** | arrival\_date\_day\_of \_month | מספר שלם | רציף | יום ההגעה בחודש הרלוונטי  (1-31) |
| **מסביר X4** | stays\_in\_week\_nights | מספר שלם | רציף | מספר לילות השבוע (שני עד שישי) שהאורח שהה או הזמין לשהות במלון |
| **מסביר X5** | adults | מספר שלם | רציף | מספר המבוגרים בחדר |
| **מסביר X6** | meal | טקסטואלי | קטגוריאלי | BB – Bed & Breakfast  לינה וארוחת בוקר HB - Half Board  לינה וחצי פנסיון  (ארוחת בוקר וערב) FB - Full Board  לינה ופנסיון מלא (ארוחת בוקר צהרים וערב) |
| **מסביר X7** | country | טקסטואלי | קטגוריאלי | ארץ מוצא |
| **מסביר X8** | reserved\_room\_type | טקסטואלי | קטגוריאלי | קוד סוג החדר. קוד מוצג במקום ייעוד מטעמי אנונימיות |
| **מסביר X9** | total\_of\_special\_requests | מספר שלם | רציף | מספר הבקשות המיוחדות של הלקוח |

# 1. תקציר מנהלים:

מטרת חלק ב של הפרויקט היא בניית מודל רגרסיה לינארית משופר ומתאים ככל שניתן לחיזוי התעריף היומי הממוצע לחופשה במלון בהינתן נתונים שונים של מגוון פרמטרים תוך שימוש בכלים שנלמדו בקורס ותוך שימוש בתוכנת R-Studio. במהלך הפרויקט ביצענו ניתוח סטטיסטי של נתונים המבוססים על מדגם מהאוכלוסייה ובחנו את הקשר בינם לבין התעריף היומי הממוצע לחופשה במלון.

בשלב הראשון בחנו הוצאה של משתנים מהמודל בהתאם למידת השפעתם על המשתנה המוסבר שלנו - תעריף יומי ממוצע לחופשה במלון. עבור משתנים רציפים השתמשנו במקדם המתאם של פירסון ובתרשימי פיזור, בבדיקת המשתנים הקטגוריאליים השתמשנו בתרשימי BoxPlot ובחרנו לאחד קטגוריות במשתנה X7 (country)בין הקטגוריות IRL ו GBR ובמשתנה X6 (Meal) בין הקטגוריות FB ו HB. התאמנו למשתנים הקטגוריאליים משתני דמה ובחנו באמצעות גרפים משתני אינטראקציה שחשבנו שיכולים לתרום למודל שלנו. לבסוף קיבלנו את המודל הראשוני שלנו.

על מנת לשפר את המודל, השתמשנו באלגוריתמים של רגרסיה לפנים, רגרסיה לאחור ורגרסיה בצעדים לפי מדד AIC ולפי מדד BIC ובחרנו את המודל שהניב לנו את מדד ה AIC המינימלי ו המקסימלי מבין המודלים שהתקבלו והמודל המלא. קיבלנו את אותו המודל בכל השיטות שנמדדו לפי מדד ה AIC ולו ערך ה AIC הנמוך ביותר, עם זאת מדד ה שלו קטן באופן מזערי מהמודל המלא ולמרות זאת מצאנו לנכון להמשיך עם המודל שמתקבל מהרצת האלגוריתמים לפי מדד AIC.

בשלב הבא בחנו האם המודל מקיים את הנחת שוויון שונויות, הנחת הלינאריות והנחת הנורמאליות באמצעות תרשימים מתאימים ואיששנו את המסקנות באמצעות מבחנים סטטיסטים. קיבלנו שהמודל שלנו לא מקיים את הנחת הנורמאליות והנחת שוויון השונויות וניסינו לשפר אותו. השתמשנו בטרנספורמציית BoxCox על המשתנה המוסבר וקיבלנו שעבור המודל המשופר, הנחת שוויון השונויות והנחת הלינאריות כן מתקיימות אך הנחת הנורמאליות לא מתקיימת. בדקנו מודלים נוספים לשיפור ובחנו טרנספורמציה לוגריתמית ופולינומיאלית מסדר שני על משתנים מסבירים. קיבלנו שהמודל שהתקבל לאחר טרנספורמציה פולינומיאלית משפר את מדד ה ומקיים את הנחות שוויון השונויות והלינאריות ולכן החלטנו שהוא יהיה המודל הסופי המתאים ביותר לחיזוי התעריף היומי הממוצע לחופשה במלון.

# 2. עיבוד מקדים:

## 2.1 הסרה של משתנים:

על מנת לבחון האם יש להסיר חלק מהמשתנים במודל נשתמש במקדם המתאם של פירסון ובמובהקות תוצאה P-Value עבור משתנים רציפים. עבור משתנים קטגוריאליים נשתמש במובהקות תוצאה P-Value ובתרשימי Boxplot. בנוסף נרצה לבחון הסרה של משתנה מסביר אם קיים קשר לינארי חזק בינו לבין משתנה מסביר נוסף כיוון שנרצה עבור כל משתנה במודל לבחון את ההשפעה הייחודית שלו על המשתנה המוסבר.  
מקדם המתאם של פירסון עוזר לבחון את הקשר הלינארי בין 2 משתנים רציפים. כאשר ערכו מתקרב ל-1 בערך מוחלט ניתן לומר שיש קשר לינארי חזק בין המשתנים, כאשר ערכו מתקרב לאפס ניתן לומר שהקשר חלש בין המשתנים.   
מובהקות התוצאה P-Value הינה ההסתברות לקבל תוצאה זהה או קיצונית לפחות כמו זו שהתקבלה בניסוי תחת השערת האפס (אשר מחזקת את ההשערה האלטרנטיבית). במודל הרגרסיה המרובה נבדוק באיזה מידה המשתנה שנבחן מסביר את המוסבר.

נתבונן בגרף הבא המייצג את הקורלציה בין המשתנים המסבירים למשתנים המסבירים ולמשתנה המוסבר (adr) :

Chart, treemap chart

Description automatically generated

Chart, scatter chart

Description automatically generatedהמשתנה המוסבר (**adr**) אל מול (**X1 - Lead Time**) :

Text

Description automatically generatedבין המשתנה המוסבר למשתנה זה יש קורלציה 0.3235- המעידה על קשר לינארי שלילי יחסית חלש על פי מדד פירסון אך ביחס לשאר המשתנים במודל שלנו זאת הקורלציה החזקה ביותר ביחס למשתנה המוסבר. מתוך הגרף קשה לראות האם המתאם בין המשתנים טוב. בהסתכלות על מודל הרגרסיה התקבל 0.000233=P-Value נמוך מאוד וזה מחזק את הטענה שיתכן ויש קשר בין המשתנים. ציפינו שיהיה קשר בין מספר הימים מתאריך ההזמנה לתאריך ההגעה (X1) לתעריף היומי הממוצע (Y) כי יתכן שהביקוש לתאריכים רחוקים יותר יהיה נמוך בהשוואה לתאריכים הקרובים יותר למועד ההגעה ובהתאם לרמת הביקוש המחיר יקבע וישפיע על תעריף הממוצע היומי. **לכן נבחר** שלא להסיר **את משתנה זה מהמודל.**

Chart, scatter chart

Description automatically generatedהמשתנה המוסבר (**adr**) אל מול (**X3 - Arrival Date Day Of Month**):

Text

Description automatically generatedבין המשתנה המוסבר למשתנה זה יש קורלציה נמוכה מאוד 0.0309. מכך ומתוך הגרף לא ניתן לזהות מתאם בין המשתנים. בהסתכלות על מודל הרגרסיה התקבל 0.733=P-Value גבוה. יש חודשים שתחילת החודש יהיה יותר מבוקש ויש כאלו שבסופו והסתכלות על הימים בלבד מתעלמת מהמידע החשוב שהוא החודש עצמו ולא מסבירה נכון את המשתנה המוסבר. **לכן נבחר** להסיר **את משתנה זה** מהמודל כיוון שהנתונים מעידים על כך שאינו מסביר טוב את המשתנה המוסבר ואין קשר סיבתי ניכר.

Chart, scatter chart

Description automatically generatedהמשתנה המוסבר (**adr**) אל מול (**X4 - Stays In Week Nights**) :

Text

Description automatically generatedבחלק א' של הפרויקט ביצענו הוצאת חריגים ממשתנה זה וכעת נבחן את ההשפעה שלו על המשתנה המוסבר בהיעדר התצפיות החריגות. כעת בין המשתנה המוסבר למשתנה זה יש קורלציה 0.1408- המעידה על קשר לינארי שלילי חלש מאוד. מכך ומתוך הגרף לא ניתן לזהות מתאם בין המשתנים. בהסתכלות על מודל הרגרסיה התקבל 0.354=P-Value גבוה שמחזק את הטענה שמשתנה זה לא מסביר בצורה טובה את המשתנה המוסבר ולדעתנו קשה להבחין בקשר סיבתי בין מספר לילות השבוע (שני עד שישי) X4 לבין התעריף היומי הממוצע (Y). **לכן נבחר** להסיר **את משתנה זה מהמודל**.

Chart, line chart

Description automatically generatedהמשתנה המוסבר (**adr**) אל מול (**X5 - Adults**):

Text

Description automatically generatedבין המשתנה המוסבר למשתנה זה יש קורלציה 0.1989 המעידה על קשר לינארי חלש על פי מדד פירסון אך ביחס לשאר המשתנים במודל שלנו זהו המתאם השני בחוזקתו מול המשתנה המוסבר. מהגרף ניתן לראות שככל שמספר המבוגרים גדל, כך תעריף הממוצע היומי עולה המעיד על קשר ליניארי חיובי. בהסתכלות על מודל הרגרסיה התקבל 0.0261=P-Value נמוך וזה מחזק את הטענה שיתכן ויש קשר בין המשתנים. ציפינו שיהיה קשר בין מספר המבוגרים בחדר(X5) לתעריף היומי הממוצע(Y) כיוון שככל שיש יותר מבוגרים בחדר המחיר עולה לאותו מספר ימים וכך גם משפיע על התעריף היומי הממוצע. **לכן נבחר** שלאלהסיר **את משתנה זה מהמודל**.

Chart, scatter chart

Description automatically generatedהמשתנה המוסבר (**adr**) אל מול (**X9 - Total Of Special Requests**) :

Text

Description automatically generatedבין המשתנה המוסבר למשתנה זה יש קורלציה 0.0822 המעידה על קשר לינארי חלש מאוד. מתוך הגרף לא ניתן לזהות מתאם בין המשתנים. בהסתכלות על מודל הרגרסיה התקבל 0.362=P-Value גבוה שמחזק את הטענה שמשתנה זה לא מסביר בצורה טובה את המשתנה המוסבר ולא ניתן לראות קשר סיבתי בין מספר הבקשות המיוחדות של הלקוח לתעריף הממוצע היומי. **לכן נבחר** להסיר **את משתנה זה מהמודל**.

משתנים קטגוריאליים:

המשתנה המוסבר (**adr**) אל מול (**X2 - Arrival Date Month**) :

Chart, scatter chart

Description automatically generatedניתן לראות מתרשים ה Boxplot שוני במדד הממוצע והחציון בין החודשים השונים ביחס למשתנה המוסבר. בנוסף ערך ה P-Value נמוך מאוד ומעיד על כך שהמשתנה מסביר בצורה טובה את המשתנה המוסבר. הנתונים האלו מסתדרים עם ההבנה שבחודשים שונים יש ביקושים ותמחור שונה של החופשה במלון ומכך שחודש החופשה משפיע על התעריף היומי הממוצע. **לכן נבחר** שלא להסיר **את משתנה זה מהמודל.**

Text

Description automatically generated

המשתנה המוסבר (**adr**) אל מול (**X6** **-** **Meal**) :

ניתן לראות מתרשים ה - BoxPlot שאין הבדלים משמעותיים בערך החציון והממוצע בין סוגי הפנסיון השונים ובנוסף ערך ה P-Value גבוה מאוד ומעיד על כך שהמשתנה לא מסביר בצורה טובה את המשתנה המוסבר. למרות שתרשים הפיזור לא מראה קשר מובהק בין המשתנה הזה למשתנה המוסבר, אנחנו חושבים שלסוג הפנסיון יש השפעה על התעריף היומי הממוצע כי כל סוג פנסיון מתומחר שונה, עם זאת נרצה לאחד בין סוג פנסיון FB ו HB כיוון שעבורם הממוצע והחציון דומים במיוחד ותרומתם למשתנה המוסבר דומה. **לכן נבחר** שלא להסיר **את משתנה זה מהמודל.**

Text

Description automatically generatedChart, box and whisker chart

Description automatically generated

המשתנה המוסבר (**adr**) אל מול (**X7 - Country**) :

ניתן לראות מתרשים ה Boxplot שאין הבדלים משמעותיים בערך החציון והממוצע בין המדינות השונות מלבד ספרד(ESP) אך ערך ה P-Value נמוך מאוד ומעיד על כך שהמשתנה מסביר בצורה טובה את המשתנה המוסבר. למרות שתרשים הפיזור לא מראה קשר מובהק בין המשתנה הזה למשתנה המוסבר, אנחנו חושבים שלמדינה יש השפעה על התעריף היומי הממוצע כי לכל מדינה יש כלכלה שונה ותמחור שונה של החופשה, עם זאת נרצה לאחד בין סוגי המדינות אירלנד (IRL) ובריטניה (GBR) כיוון שלהן אופי ואורך החיים דומה ותרומתם למשתנה המוסבר תהיה דומה. **לכן נבחר** שלא להסיר **את משתנה זה מהמודל.**

Text

Description automatically generated**Chart, box and whisker chart

Description automatically generated**

המשתנה המוסבר (**adr**) אל מול (**X8 - Reserved Room Type**) :

Chart, box and whisker chart

Description automatically generatedText

Description automatically generatedניתן לראות מתרשים ה Boxplot שוני במדד הממוצע והחציון בין סוגי החדרים השונים ביחס למשתנה המוסבר. בנוסף ערך ה P-Value נמוך מאוד ומעיד על כך שהמשתנה מסביר בצורה טובה את המשתנה המוסבר. הנתונים האלו מסתדרים עם ההבנה שעבור חדרים שונים התמחור שונה ומשפיע על התעריף היומי הממוצע ו**לכן נבחר** שלא להסיר **את משתנה זה מהמודל.**

## 2.2 התאמת משתנים:

Chart, box and whisker chart

Description automatically generated**עבור המשתנה X7 המייצג את המדינות** **בהן מתבצעת החופשה**, ראינו לנכון לאחד את המדינה אירלנד (IRL) עם המדינה בריטניה (GBR). ניתן לראות בתרשים ה Boxplot בסעיף 2.1 שהממוצע והחציון בשתי מדינות דומה מאוד, בנוסף לאירלנד יש 3 תצפיות בסה"כ בעלות השפעה זניחה על המודל. מכאן ומתוך ההבנה שבריטניה ואירלנד הן מדינות דומות באופיין התיירותי עקב מיקומן הגיאוגרפי ואורך החיים הדומה בשתי המדינות אנחנו חושבים שלשתיהן יש השפעה דומה על המשתנה המוסבר ונרצה לאחד אותן לקטגוריה אחת. בתרשים ה Boxplot ראינו שהממוצע והחציון של מדינת פורטוגל דומה מאוד לבריטניה ולאירלנד אך בעקבות השוני במיקום הגיאוגרפי ובתרבות בין המדינות, איחוד זה אינו נכון מבחינה רעיונית ונבחר שלא לחבר בין פורטוגל בריטניה ואירלנד. מצורף תרשים Boxplot לאחר איחוד הקטגוריות.

Chart, box and whisker chart

Description automatically generated

**עבור המשתנה X6 המייצג את סוג הפנסיון**, ראינו לנכון לאחד את סוג הפנסיון FB עם סוג פנסיון HB. ניתן לראות בתרשים ה Boxplot בסעיף 2.1 שהממוצע והחציון עבור שתי הקטגוריות דומה מאוד. בנוסף, ישנם 4 תצפיות בלבד לסוג פנסיון FB אשר בעלות השפעה זניחה על המודל. מכאן ומתוך ההבנה שההבדל בין בחירת FB ל HB לא משמעותית לעומת בחירת פנסיון BB שבו יש רק ארוחת בוקר שהיא הארוחה הכי קלה מבין כל הארוחות וסביר להניח שהיא חבילת הבסיס בהזמנת חופשה (ללא עלות נוספת לעומת FB ו HB). אנחנו חושבים שלקטוגריות FB ו HB יש השפעה דומה על המשתנה המוסבר ונרצה לאחד אותן לקטגוריה אחת. מצורף תרשים Boxplot לאחר איחוד הקטגוריות.

## 2.3 הגדרת משתנה דמה:

נגדיר משתני דמה עבור המשתנים הקטגוריאליים במודל:

עבור משתנה **X6 - Meal:**קבוצת הבסיס תהיה BB ונגדיר משתנה דמה אחד לFB/HB :

עבורמשתנה **X2 - Arrival Date Month:**

קבוצת הבסיס תהיה August וניצור 8 משתני דמה:

עבור משתנה **X7 - Country:**

קבוצת הבסיס תהיה PRT וניצור 3 משתני דמה:

**עבור משתנה X8 - Reserved Room Type:**

קבוצת הבסיס תהיה A וניצור 3 משתני דמה:

## 2.4 משתני אינטראקציה:

**משתנה אינטראקציה בין X5 - Adults לבין X8 - Reserved Room Type:**

Chart, line chart

Description automatically generatedבחרנו לבחון משתנה אינטראקציה זה כיוון שסביר להניח שסוגי חדרים שונים יתאימו למספר שונה של מבוגרים. מהגרף ניתן לראות שההבדל בין השיפועים לא משמעותי עבור חדרים מסוג A, E וF . עבור חדר מסוג D יש הבדל בשיפוע ביחס לשאר סוגי החדרים אך בגלל שערך ה P-Value גבוה אנחנו נבחר שלא להוסיף את המשתנה הזה למודל שלנו.

Text

Description automatically generated

**משתנה אינטראקציה בין X1 - Lead Time לבין X7 - Country:**

בחרנו לבחון משתנה אינטראקציה זה כיוון שסביר להניח שבין המדינות יש הבדל בתיירות ובביקוש וביעדים בעלי ביקוש גבוה יש צורך לבצע הזמנה של החופשה מוקדם יותר מהיעדים עם ביקוש נמוך. מהגרף ניתן לראות כי קיים הבדל בין השיפועים אך משתנה האינטראקציה לא מובהק ולכן לא נוסיף את משתנה זה למודל.

Chart, line chart

Description automatically generatedA black screen with white text

Description automatically generated with low confidence

**משתנה אינטראקציה בין X1 - Lead Time לבין X8 - Reserved Room Type:**

בחרנו לבחון משתנה אינטראקציה זה כיוון שישנם חדרים בעלי ביקוש גבוה שיש צורך להזמין אותם תקופה ארוכה יותר מראש על מנת לשריין בהם מקום ולדעתנו סוג החדר משפיע על ה Lead Time.

מהגרף ניתן לראות כי בין חדרים מסוג D ו E השיפוע כמעט זהה, אך השיפוע של חדר מסוג F ו A שונה.

A black screen with white text

Description automatically generated with low confidenceChart

Description automatically generatedמהסתכלות על מודל הרגרסיה המשתנים האינטראקציה של החדרים D ו E אינם מובהקים אך רמת המובהקות של חדר D יחסית נמוכה, ולכן נבחר להכניס את משתנה האינטראקציה הזה למודל שלנו.

# 3. התאמת המודל ובדיקת הנחות המודל:

## 3.1 בחירת משתני המודל:

על מנת לבחון ולבחור את משתני המודל שלנו נשתמש באלגוריתמים של רגרסיה לאחור, רגרסיה לפנים ורגרסיה בצעדים לפי מדד AIC ולפי מדד BIC. את המודל נבחר לפי ערך הAIC המינימלי המתקבל. במצב של שוויון במדד הAIC נבחר את המודל שיניב הגדול ביותר.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| BIC Stepwise | AIC Stepwise | BIC Backward | AIC Backward | BIC Forward | AIC Forward | Full Model |  |
| 763.774 | 759.444 | 763.774 | 759.444 | 763.774 | 759.444 | 761.380 | AIC |
| 809.0272 | 810.354 | 809.0272 | 810.354 | 809.0272 | 810.354 | 820.775 | BIC |
| 0.7708 | 0.7816 | 0.7708 | 0.7816 | 0.7708 | 0.7816 | 0.7825 |  |

קיבלנו את ערך ה AIC המינימלי מכל אחד מהאלגוריתמים המנוחים לפי מדד ה AIC. למרות שבמודל המלא שלנו קיבלנו ערך גבוה יותר בפער יחסית מזערי (0.0009) ,אנחנו נבחר את המודל לפי מדד ה AIC מכוון שבמודל שהתקבל לפי AIC הוסר רק משתנה האינטראקציה שהוספנו בסעיף הקודם. בנוסף ניתן לראות שבכל אחד מן האלגוריתמים משתנה האינטראקציה הוסר ואף ב AIC Backward הוא המשתנה שהוסר ראשון. מוטיבציה נוספת לקחת את המודל ללא משתנה האינטראקציה היא שבסעיף הקודם אכן ראינו שהוא אינו מובהק לחלוטין.

**המודל המלא:**

adr ~ lead\_time + adults +

factor(arrival\_date\_month) +

factor(country) +

factor(reserved\_room\_type) +

factor(meal) +

lead\_time \* reserved\_room\_type

**המודל הנבחר:**

adr ~ lead\_time + adults +

factor(arrival\_date\_month) +

factor(country) +

factor(reserved\_room\_type) +

factor(meal)

## 3.2 בדיקת הנחות המודל:

בדיקת הנחת הנורמאליות של השגיאות:

בהסתכלות על ההיסטוגרמה ועל גרף Normal Q-Q ,המתאר את ההתנהגות המצופה מנתונים המגיעים מהתפלגות נורמלית, ניתן לראות שרוב התצפיות לא נמצאות על הקו הישר, כלומר שההתפלגות לא נורמלית. נבחן את ההנחה באמצעות מבחנים סטטיסטים.

Chart, line chart

Description automatically generatedChart, histogram

Description automatically generated

מבחנים סטטיסטיים לבדיקת קיום הנחת הנורמאליות :

Text

Description automatically generatedמבחן Kolmogorov – Smirnov :

בהרצת מבחן זה קיבלנו P-Value קטן מאוד ולכן נדחה את השערת האפס ברמת מובהקות 5% **ונאמר שהנחת הנורמאליות לא מתקיימת במודל זה**. תוצאות המבחן מחזקות את מה שראינו מהתרשימים.

מבחן Shapiro – Wilk:

**Text

Description automatically generated**גם בהרצת מבחן זה, שנחשב לחזק יותר מהמבחן הקודם, קיבלנו P-Value קטן מאוד ולכן נדחה את השערת האפס ברמת מובהקות 5% **ונאמר שהנחת הנורמאליות לא מתקיימת במודל זה.**

בדיקת הנחת הלינאריות:

בהסתכלות על תרשים פיזור השגיאות המתוקננות לעומת ערך החיזוי ניתן לראות פיזור יחסית אחיד סביב האפס, לכן נסיק כי הנחת הלינאריות מתקיימת ונבחן זאת באמצעות מבחן סטטיסטי Chow.

Chart, scatter chart

Description automatically generated

Text

Description automatically generatedמבחן Chow לבחינת הנחת הלינאריות:  
בהרצת מבחן Chow קיבלנו P-Value= 0.4942 ולכן נחליט שלא לדחות את השערת האפס ברמת מובהקות 5% ו**נאמר שהנחת הלינאריות מתקיימת.** תוצאות המבחן מחזקות את מה שראינו מהגרף.

בדיקת הנחת שוויון שונויות:

Chart, scatter chart

Description automatically generatedנבדוק את הנחת שוויון השונויות באמצעות תרשים פיזור השגיאות המתוקננות לעומת ערך החיזוי. על פי הגרף ניתן לראות שעבור ערכים גבוהים יותר מתקבלת שונות שגיאות גבוהה יותר. קיבלנו מעין צורה של משפך הגדל מצד שמאל לימין, לכן על פי הגרף נאמר ששוויון השונויות לא מתקיים במודל זה.

מבחן Goldfeld - Quandt לבחינת הנחת שוויון השונויות:

לצורך ביצוע המבחן חיפשנו משתנה במודל שחשדנו בו שיתכן שגורם לאי שוויון השונויות ומצאנו את המשתנה LeadTime כחשוד. בהסתכלות על הגרף ניתן לראות פיזור בצורת משפך הקטן משמאל לימין ומראה על אי שוויון שונויות.

ביצענו בדיקה של השונויות בתחומים שונים למשתנה זה וקיבלנו:

Text

Description automatically generated with medium confidence

התוצאות שקיבלנו מתיישבות עם הגרף.

תוצאות מבחן Goldfeld - Quandt :

קיבלנו P-Value נמוך מאוד ולכן נדחה את השערת האפס ברמת מובהקות 5% ונאמר **אין במודל שוויון שונויות**.

Chart, scatter chart

Description automatically generated**Text

Description automatically generated**

# 4. שיפור המודל:

Chart, diagram

Description automatically generatedבבדיקות שביצענו על המודל קיבלנו שהנחת שוויון השונויות והנחת הנורמאליות לא מתקיימות. נבצע מבחן BoxCox כדי לבחון איזה טרנספורמציה לבצע על המשתנה המוסבר .(adr) מהמבחן ניתן לראות שאפס נמצא ברווח הסמך ומכאן נבחר לבצע טרנספורמציה עבור , כלומר .

Text, letter

Description automatically generated

לאחר בדיקה עבור מודל ה , נראה שבעבור מודל זה הנחות שוויון שונויות והנחת הליניאריות כן מתקיימות, אך הנחת הנורמליות של השגיאות אינו מתקיים. בדקנו מודלים נוספים וראינו שלמרות שיש מודלים בעלי R2adj גבוה יותר ממודל , פחות הנחות מתקיימות במודלים האחרים ולכן נבחר להישאר עם מודל זה.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| שוויון שונויות | ליניאריות | נורמליות | R2adj | טרנספורמציה |
| לא מתקיים | מתקיים | לא מתקיים | 0.7816 |  |
| מתקיים | מתקיים | לא מתקיים | 0.7881 |  |
| לא מתקיים | מתקיים | לא מתקיים | 0.706 |  |
| לא מתקיים | מתקיים | לא מתקיים | 0.7962 |  |
| מתקיים | לא מתקיים | לא מתקיים | 0.7598 |  |

לאחר שבחרנו במודל נרצה לבדוק האם טרנספורמציה על המשתנים המסבירים יכולה לתרום לקיום ההנחות הנדרשות במודל. נבצע טרנספורמציה לוגריתמית ופולינומיאלית מסדר 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| שוויון שונויות | ליניאריות | נורמליות | R2adj | טרנספורמציה |
| לא מתקיים | מתקיים | לא מתקיים | 0.7689 | Log |
| מתקיים | מתקיים | לא מתקיים | 0.7883 | X2 |

נראה שבעבור המודל עם הטרנספורמציה פולינומיאלית מסדר שני אנחנו מקבלים שהנחת הליניאריות והנחת שוויון השונויות מתקיימות בניגוד למודל הטרנספורמציה הלוגריתמית, בנוסף נוכל לראות שמדד ה R2adj גדל ולכן אנחנו נבחר במודל זה להיות המודל הסופי.

\* בחרנו שלא להחזיר למודל משתנים שהוסרו קודם כיוון שהמודל לאחר הסרתם מספק נתונים מובהקים יותר ועם קשר סטטיסטי חזק יותר.

המודל הסופי:

Log(adr) ~ (lead\_time)2 **+ (**adults)2 **+**

factor(arrival\_date\_month) +

factor(country) **+**

factor(reserved\_room\_type) +

factor(meal)

([קישור לבדיקות בנספחים](#_בדיקת_הנחות_מודלים)).

# נספחים

## אינטראקציות נוספות שנבדקו :

X5 - Number of Adults & X2 - Months.

Chart, line chart

Description automatically generated

X5 - Number of Adults & X6 - Meal.

Chart, line chart

Description automatically generated

X5 - Number of Adults & X7 - Countries.

Chart

Description automatically generated

X1 - Lead Time & X2 - Months.

Chart

Description automatically generated

X1 - Lead Time & X6 - Meal.

Chart, line chart

Description automatically generated

## בניית המודל :

Full Model

Text

Description automatically generated

Empty Model

Text

Description automatically generated

AIC – FORWARD

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

BIC – FORWD

Text

Description automatically generated

Text

Description automatically generated

AIC – BACK

Text

Description automatically generated

Text

Description automatically generated

BIC – BACK

Text

Description automatically generated

Text

Description automatically generated

AIC – STEPS

Text

Description automatically generated with medium confidence

Text

Description automatically generated

Text

Description automatically generated

BIC – STEPS

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

## בדיקת הנחות מודלים לשיפור :

ערך Lambda:  
Text

Description automatically generated

הנחות מודל עבור - Log( Y )

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, histogram

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

# הנחת הליניאריות ושוויון השונויות מתקיימת.

הנחות המודל עבור -

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, histogram

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

# רק הנחת הליניאריות מתקיימת.

לאחר שנבצע טרנספורמציית שורש עבור המשתנים Lead Time ו Adults :

Text

Description automatically generated

הנחות המודל עבור -

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, histogram

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

# רק הנחת הליניאריות מתקיימת.

הנחות המודל עבור –

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated Chart, histogram

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

# לא עומד בהנחות המודל מלבד שוויון שונויות.

בדיקת הנחות למודל Log עם טרנספורמציה למשתנים מסבירים בLog :

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated Chart, histogram

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated  
Text

Description automatically generated

בדיקת הנחות למודל ln(Y) עם טרנספורמציה פולינומיאלית למשתנים מסבירים בסדר שני :

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated Chart, histogram

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated