Unsupervised learning

Diabetic Data analysis

By: Matan Vetzler 314883943 & Rachel Portnoy 208938464

תוכן עניינים:

הקדמה

הדאטא הנ"ל מכיל רשומות מבתי חולים ברחבי ארצות הברית. הדאטא כולל מידע אודות 101767 מפגשים של מטופלים שונים בעלי רקע סוכרתי, אשר תיעדו את מהלך ביקורם בבית החולים ואת המידע הנאסף מהם במהלך הביקור.

המידע במאגר נתונים זה נאסף במהלך 10 שנים 1999-2008 מ130 בתי חולים בגדלים שונים ברחבי ארצות הברית. מאגר הנתונים מכיל 55 מאפיינים (פיצ'רים) המתארים את המטופל והמפגש נתונים אלו מכילים נתונים דמוגרפיים, מידע על תרופות ,אישפוזים והמסקנות אשר אבוחנו במהלך שהותם.

מאגר זה מכיל נתונים בעלי חשיבות גבוהה אך ישנו קושי בעבודה עם נתונים אלה מפני שהם הטרוגניים, בלתי עקביים, בעלי ערכים חסרים, בעלי מימדים שונים לא רק במספר התכונות אלא גם במורכבות שלהם. בנוסף ניתוח של מאגרי נתונים הוא קשה יותר מאשר ניתוח תוצאות ניסוי מבוקר שכן אין השפעה על אופן איסוף המידע.

במסמך זה נציג את תהליך העבודה, התוצאות והמסקנות העולות מניתוח המידע ושימוש בשיטות שונות של למידה לא מפוקחת אשר נלמדו במהלך הקורס.

שאלות מחקר:

שאלת המחקר העיקרית שהובילה אותנו בתהליך הניתוח הינה

האם ניתן לנבא אישפוז של מטופל בתקופה קצרה של 30 יום מאז הביקור האחרון שלו ע"פ הנתונים המופיעים ברשומות.

מעבר, נרצה להסיק כל מידע שימושי אשר מביא לתובנות חדשות ומעניינות מתוך הדאטא המתועד.

תהליך הניתוח כלל מספרים שלבים אותם נתאר בצורה מורחבת בהמשך המסמך:

* ניקוי המידע על מנת שהעבודה תהיה קלה ונוחה יותר.

*ניתוח ראשוני של הדאטא באמצעות היסטוגרמות, קורלציות וניתוחים סטטיסטיים.

*צמצום הדטה על פי הניתוח הראשוני וניתוח מעמיק יותר על מספר מצומצם של תכונות (פיצ'רים) .

*הפעלת אלגוריתמים ללמידה לא מפוקחת שנלמדו בכיתה.

טיפול ועיבוד המידע הראשוני

טיפול ראשוני בדאטא מהווה חלק אינטגרלי בתהליך עבודה בלמידת מכונה. היות והאיכות של הדאטא והאינפורמציה אשר אנו יכולים להסיק ממנה משפיע ישירות על איכות תוצאות המודל\קלאסטריזציה שלנו. על כן, זה חשוב מאוד נשעבד את הדאטא לפני תהליך של למידת מכונה ושימוש בו.

בעבודה שלנו, ישנם שלושה חלקים דומיננטיים בטיפול ועיבוד ראשוני של הדאטא. השלבים הינם:

1. מחיקת דאטא אשר לא מספק ו\או לא יכול לספק לנו אפשרות ללמוד ממנו כראוי או להסיק ממנו מסקנות בצורה טובה מספיק.

2. העברת המידע הקטגורי, בין אם היינו nominal \ ordinal לכדי ייצוג מספרי בכדי שנוכל להסיק ממנו מסקנות וללמוד ממנו. היות וכלי למידה וצורת הסקת מסקנות מן הדאטא אשר אנו משתמשים בהם דורשים דאטא אשר מיוצג בצורה מספרית.

3. נירמול המידע המספרי לטווח מספרי מסוים בכדי "לעזור" לאלגוריתמי הלמידה והקלאסטריזציה להתכנס בצורה מהירה יותר, ובכדי להסיק מסקנות בצורה נוחה יותר כאשר הדאטא מסודר בקנה מידה נוח.

במהלך הסקשן הבא, נתאר בצורה פרטנית את תהליך הטיפול והעיבוד הראשוני בדאטא ה"נא" לכדי דאטא מוגמר ומוכן לתהליך הסקה, למידה וקלסטריזציה, בכדי להציג את הטיפול הספציפי בסט המידע הסוכרתי, אשר עליו אנו נבחנים בעבודת גמר זה.

מחיקת מידע לא אינפורמטיבי

כאמור, כחלק מתהליך העיבוד הראשוני של הדאטא, ישנו צורך במחיקת דאטא אשר אינו מספק לנו ו\או אינו מסוגל לספק לנו יכולת הסקה מספקת מתוכו את המידע אותו אנחנו מחפשים לחקור.
על כן, העברנו את קובץ המידע ה"נא" שלנו – diabetic_data.csv בpipeline אשר מהווה תהליך מאחד מחיקת מידע ספציפי לקראת למידה וקלסטריזציה, אשר נתאר עליו בהמשך בפירוט, כמו שאפשר לראות בcode snippet הבא:

תהליך המחיקה הנ"ל מחולק ל5 מתודות, אשר כל מתודה אחראית לצורת מחיקה אחרת אשר פועלת על סוג שונה של מידע.

```
class DataEraser:
    def __init__(self, diabetic_data):
        self.diabetic_data = diabetic_data

    def delete_unnecessary_data(self):
        """
        function uniting all deletion manipulation techniques on the diabetic data
        :return: class object with diabetic data after applying deletion methods on
        """
        return self._delete_duplicated_patient_nbr() \
            ._delete_uninformative_codes_data() \
            ._delete_gender_column() \
            ._delete_high_percentage_missing_data_columns() \
            ._delete_missing_data_rows() \
            ._delete_specific_medications_columns() \
            .diabetic_data
```

המתודה הראשונה היינה המתודה בעלת החתימה delete_duplicated_patient_nbr, אשר מוצגת להלן:

```
def _delete_duplicated_patient_nbr(self):
    """
    deleting duplicated same patient inpatient/outpatient occurrence
    :return: class object with diabetic data after patient number duplicate deletion
    """
    self.diabetic_data.drop_duplicates(subset='patient_nbr', inplace=True)

    return self
```

מתודה זו אחראית כהתחלה, למחיקת רשומות אשר מהוות ביקורים חוזרים של אותו החולה אשר מסתכמות בכ30 אלף רשומות. החלטנו לפעול בצורה זו בכדי שכל רשומה תהי בלתי תלויה באחרת, בצורה זו נוכל ללמוד באופן קל ופשוט יותר את המסקנות אשר דאטא זה מספק לנו.

המתודה הבאה היינה המתודה בעלת החתימה – delete_uninformative_codes_data אשר מוצגת להלן:

```
def _delete_uninformative_codes_data(self):
    """
    deleting columns with id information which don't provide useful
    information clustering wise
    :return: class object with diabetic data after uninformative id data deletion
    """

    # deleting unique elements column - encounter id and patient number, as don't provide useful
    # information clustering wise
    self.diabetic_data.drop(['encounter_id', 'patient_nbr'], axis=1, inplace=True)

    # deleting discharge disposition id elements which relates for death or hospice
    # as described at IDs_mapping.csv file, since those patients won't be able to be readmitted
    death_hospice_codes = [11, 13, 14, 19, 20, 21]
    self.diabetic_data = \
        self.diabetic_data.loc[~self.diabetic_data.discharge_disposition_id.isin(death_hospice_codes)]

    return self
```

מתודה זו אחראית למחיקת מידע אשר מיוצג בצורה "קודית" אשר לא מהווה מידע שימושי.

כהתחלה, מתודה זו מוחקת את העמודות 'patient_nbr' & 'encounter_id' מאחר ועמודות אלה מהוות מידע uniqueי אשר לא מהווה שום מידע שימושי לתהליך ההסקה והלמידה שלנו.

בנוסף, מתודה זו מוחקת מעמודת 'discharge_disposition_id' אשר הערכים הספצפיים מתוך הסט – [11, 13, 14, 19, 20, 21] אשר קודים אלו על פי המיפוי מייצגים שחרור מסיבת מוות או מסיבת שחרור להוספיס, מה שמוביל לכך שאותו חולה לא יוכל לחזור אל בית החולים, מה שמגביל אותנו טוטאלית מלמידה והסקה מאותה רשומה.

המתודה הבאה היינה המתודה בעלת החתימה – delete_gender_column המופיעה להלן:

```
def _delete_gender_column(self):
    """
    deleting gender columns due to approx same number of male-female patients ration,
    with approx same percentages of readmitted levels
    :return: class object with diabetic data after gender column deletion
    """
    self.diabetic_data.drop(['gender'], axis=1, inplace=True)

    return self
```

מתודה זו אחראית למחיקת העמודה המתעדת את המידע מותאם מין.

אנו מוחקים מידע זה היות ופיזור הגברים והנשים המתועדות בסט דאטא זה היינו "חצי חצי", בנוסף אנו רואים כי אם נסתכל על החלוקה מבחינת אחוזים לשלושת ערכי עמודת 'readmitted' נראה התפלגות דומה לחלוטין בין הנשים לגברים.

מעבר לכך, נראה כי עמודה זו אינה מספקת לנו שם מידע נוסף אשר רלוונטי למחקר שלנו ועל כן בחרנו למחוק אותה.

המתודה הבאה היינה המתודה בעלת החתימה – delete_high_percentage_missing_data_columns אשר מוצגת להלן :

```
def _delete_high_percentage_missing_data_columns(self):
    """
    deleting columns with high level of missing data percentage
    :return: class object with diabetic data after high level of missing data percentage columns deletion
    """

    # weight column missing rate : 97%
    # payer_code missing rate: 52%
    self.diabetic_data.drop(['weight', 'payer_code'], axis=1, inplace=True)

    return self
```

מתודה זו אחראית למחיקת עמודות אשר חסר בהן מידע רב יחסית, אשר לא נרצה להשלימו על ידי שיטות interpolation מאחר שיעבדו בצורה לא אופטימלית ולא מדויקת מספיק מתוך חוסר רב במידע.

העמודות האלו היינן 'payer_code' & 'weight' אשר חסרות 52% ו97% מן הדאטא, בהתאמה.

יש לציין כי מן העמודה 'medical_specialty' חסרים 47% מן הדאטא, אך על עמודה זו, נפעיל טיפול שונה.

המתודה הבאה היינו המתודה בעלת החתימה – delete_specific_medications_columns אשר מוצגת להלן:

```python
def _delete_specific_medications_columns(self):
    """
    deleting specific medications columns
    :return: class object with diabetic data after specific medications columns
    """

    # the following medicines were provided to extremely small amount of hospitalized
    # patients (1 - 10), thus cannot learned from, due to low level of documentations
    # around 100K examples
    low_versatile_columns = ['acetohexamide', 'citoglipton', 'examide', 'glimepiride-pioglitazone',
                             'metformin-pioglitazone', 'metformin-rosiglitazone', 'troglitazone', 'acarbose']
    self.diabetic_data.drop(low_versatile_columns, axis=1, inplace=True)

    # the following medicines do not provide any useful or interesting information thus being deleted
    useless_columns = ['metformin', 'repaglinide', 'nateglinide', 'chlorpropamide', 'glimepiride', 'glipizide',
                       'glyburide', 'tolbutamide', 'pioglitazone', 'rosiglitazone', 'tolazamide',
                       'glyburide-metformin', 'glipizide-metformin']
    self.diabetic_data.drop(useless_columns, axis=1, inplace=True)

    return self
```

מתודה זו אחראית למחיקת עמודות אשר מייצגות מידע הקשור להגשת תרופות מסוימות לחולה.

כהתחלה, אנו מוחקים עמודות ספציפיות אלו המופיעות תחת המשתנה 'low_versatile_columns' מאחר ועל פי הבדיקה, חוץ מ1 עד 10 תיעודים של נתינת תרופות אלו, רוב התיעוד מציג כי לא הוגשו תרופות אלו לחולים, על פי כן אין לנו אפשרות ללמוד מקובץ תיעודים כה קטן.

לאחר מכן, בעקבות חקירה של שאר התרופות ואופן השפעתן, מצאנו כי כלל התרופות אשר מופיעות תחת המשתנה 'useless_columns' משפיעות בצורה מעטה ביותר ולכן נמחקו היות ולהשאיר אותן במקום למחוק אותן מפסיד מבחינת עלות מול תועלת.

לבסוף, נציג את העובדה כי התרופות 'insulin' & 'miglitol' נשארו מאחר וראינו כי נוכל להסיק מסקנות מעניינות ותורמות למחקר שלנו על פיהן.

המתודה הבאה והאחרונה היינה המתודה בעלת החתימה – delete_missing_data_rows אשר מוצגת להלן:

```
def _delete_missing_data_rows(self):
    """
    deleting lines with missing features data
    :return: class object with diabetic data after rows with missing features data deletion
    """

    # after deleting high level of missing data percentage columns
    # deleting few of the lines containing missing values represented in out
    # data set as - ["?", "Unknown/Invalid"]
    self.diabetic_data = self.diabetic_data.replace(['?', 'Unknown/Invalid'], np.nan)
    all_columns = list(self.diabetic_data.columns.values)
    all_columns.remove('medical_specialty')
    self.diabetic_data.dropna(inplace=True, subset=all_columns)

    return self
```

לאחר שמחקנו את העמודות אשר מחסירות מידע רב מכלל ערכיהן, מתודה זו אחראית למחיקת רשומות אשר מחסירות מידע, אשר בסט המידע שלנו מיוצג בסימון – ['?', 'Unkown/Invalid'].

גם כאן נעדיף למחוק רשומות אלה במקום להפעיל שיטות interpolation. יש לשים לב כי איננו מוחקים רשומות אשר מכילות מידע חסר בעמודה 'medical_specialty' היות ויש לה טיפול שונה, כמו שציינו לפני.

<u>סיכום חלק מחיקת מידע</u>

לאחר הפעלת כלל מתודות מחיקה אלו על המידע המקורי שלנו, נשארנו עם כ67 אלף רשומות על 38 עמודות (פיצ'רים)
השלב הבא היינו הצגת מידע מגוון זה, אשר מכיל ערכים מספריים, קטגורים (בין אם nominal לordinalים) לכדי ייצוג מספרי. נציג את תהליך זה בפרק הבא.

קידוד המידע לכדי מידע נומרי

לאחר השלב הראשון בו מחקנו חלק מן המידע אשר לא מהווה לנו מידע אינפורמטיבי יעיל למטרת הלמידה וההסקה שלנו, נרצה לקחת את הדאטא המעובד ולקודד את המידע האגור בו לכדי מידע נומרי בלבד, בכדי שנוכל לבצע עליו הסקה, למידה וקלסטריזציה כראוי.

כמו בשלב המחיקה, מימשנו pipeline המהווה תהליך קידוד (encoding) על הדאטא diabetic_data.csv לאחר תהליך המחיקה בפרק הקודם.

תהליך הקידוד חולק בקוד לשני חלקים, החלק הראשון היינו תהליך קידוד ידני אשר אנו בחרנו כיצד לקודד את הדאטא, לאחר חקירה מדעית אודות הפיצ'רים הקיימים בדאטא, ומישקלנו אותם בהתאם. החלק השני היינו תהליך קידוד אוטומטי על ידי כלי קידוד מידע קטגורי הקיימים בשוק.

נוכל לראות להלן את הקוד המהווה את pipeline הקידוד אשר לאחר ביצוע כלל שלבי הקידוד מחזיר את הדאטא המעובד:

```
class DataEncoder:
    def __init__(self, diabetic_data):
        self.diabetic_data = diabetic_data

    def encode_data(self):
        """
        main function responsible to encode diabetic data after erasing specific data by custom and automatic methods
        :return: data set after encoding it's data
        """
        print(self.diabetic_data)
        return self._custom_encode_categorical_data() \
            ._automatic_encode_categorical_data() \
            .diabetic_data
```

כמו שהצגנו הקוד מחולק לקידוד ידני ואוטומטי, כאשר תהליך הקידוד הידני מיוצג בצורה הבאה:

```python
def _custom_encode_categorical_data(self):
    """
    custom encoding diabetic data set
    :return: class object after custom encoding data
    """
    return self._encode_miglitol() \
        ._encode_medical_specialty() \
        ._encode_diagnosis_code() \
        ._encode_age() \
        ._encode_glucose_level() \
        ._encode_hemoglobin_level() \
        ._encode_medicines_change() \
        ._encode_nominal_data_to_str()
```

בעוד תהליך הקידוד האוטומטי מיוצג בצורה הבאה:

```python
def _automatic_encode_categorical_data(self):
    """
    automatic encoding data set with one-hot encoder and normalizing with z-score
    :return: class object after automatic encoding data
    """
    scaler = StandardScaler()
    numeric_columns = self.diabetic_data._get_numeric_data().columns
    self.diabetic_data[numeric_columns] = scaler.fit_transform(self.diabetic_data[numeric_columns])
    categorical_columns = self._get_categorical_columns()
    self.diabetic_data = pd.get_dummies(self.diabetic_data, columns=list(categorical_columns))

    return self
```

בפרק הנ״ל נציג ונסביר כל מתודת קידוד, תפקידה ומשמעותה.

נתחיל בקידוד הידני אשר מכיל 8 מתודות קידוד שונות אשר פועלות בצורה שונה, על דאטא שונה בכל פעם.

המתודה הראשונה היינה המתודה בעלת החתימה – encode_miglitol אשר מוצגת להלן:

```
def _encode_miglitol(self):
    """
    function responsible for encoding miglitol feature which stands for medicine submission
    as seen people who were submitted with lower level of miglitol surely came back.
    :return: class object with diabetic data after applying encoding methods on
    """
    miglitol_dict = {'Up': 0,
                     'Down': 500,
                     'Steady': 0,
                     'No': 0
                     }
    self.diabetic_data['miglitol'] = self.diabetic_data['miglitol'].replace(miglitol_dict)

    return self
```

מתודה זו אחראית לקודד את עמודת התרופה 'miglitol' לכדי מידע נומרי.

על פי חקירת אותה עמודה ביחס לעמודת ה'readmitted' אנו רואים כי בכל מקרה בה חולה קיבל את אותה התרופה במינון נמוך ממנו הוא רגיל, הוא בוודאות חוזר לבית החולים בתווך של 30 ימים, לכן הרגשנו כי ישנו צורך לתת משמעות כבדה לאותו פרמטר של ירידה במינון. לשאר הפרמטרים נתנו ערך אפסי מאחר ועל פי בדיקה הערכים אשר קיימים לאותה עמודת miglitol היינם הערכים של ירידה במינון או אי נתינת התרופה כלל, לכן החלטנו כי ניתן משקל לפרמטר הורדת המינון בלבד.

המתודה הבאה היינה המתודה בעלת החתימה – encode_medical_specialty בעלת החתימה הבאה:

```
def _encode_medical_specialty(self):
    """
    function responsible for encoding medical specialty column based on 10 highest informative medical specialty
    :return: class object after encoding medical specialty column
    """
    top_ten_medical_speciality = ['?', 'InternalMedicine', 'Emergency/Trauma', 'Family/GeneralPractice',
                                  'Cardiology', 'Surgery-General', 'Nephrology', 'Orthopedics',
                                  'Orthopedics-Reconstructive', 'Radiologist']

    self.diabetic_data['med_spec'] = self.diabetic_data['medical_specialty'].copy()
    self.diabetic_data.loc[~self.diabetic_data.medical_specialty.isin(
        top_ten_medical_speciality), 'med_spec'] = 'Other'
    self.diabetic_data.drop(['medical_specialty'], axis=1, inplace=True)

    return self
```

מתודה זו אחראית להכין לקידוד את עמודת 'medical_specialty'.

למרות חוסר של 47% ערכים מכלל ערכי עמודה זו, ראינו כי עמודה זו משפיעה בצורה מגוונת ומעניינת על ערכי עמודת 'readmitted' ולכן בכדי להישאר עם גודל מידע סביר שלא יגדיל לנו את מימד הפיצ'רים בצורה משמעותית, על פי כן לקחנו את עשרת ערכי התמקצעות הרפואית אשר ערכיהם חזרו בצורה הרבה ביותר, ואת כל שאר הערכים השארנו כ-Other.

מצאנו כי עשרת הערכים המובילים מבחינת שכיחות הינם המקצועות: ['InternalMedicine', 'Emergency/Trauma', 'Family/GeneralPractice', 'Cardilogy', 'Surgery-General', 'Nephrology', 'Orthopedics', 'Orthopedica-Reconstructive', 'Radiologist']

יש לציין כי אין לערכים הנ"ל משקל שונה או השפעה דומה בהכרח ביחס לשאר ההתמקצעות הרפואית, אלא לקחנו אותם היות ומכילים מספיק ערכים בכדי להסיק וללמוד מהם מה שאנחנו רוצים למצוא ולחקור.

המתודה הבאה היינה המתודה בעלת החתימה – encode_diagnosis_code אשר מוצגת להלן:

```
def _encode_diagnosis_code(self):
    """
    function responsible for encoding column - diagnosis code by each code and it's corresponding value
    :return: class object after encoding diagnosis code column
    """
    self._medically_map_diagnosis()

    # diagnosis grouping
    diag_cols = ['diag_1', 'diag_2', 'diag_3']
    for col in diag_cols:
        self.diabetic_data['temp'] = np.nan

        condition = self.diabetic_data[col] == 250
        self.diabetic_data.loc[condition, 'temp'] = 'Diabetes'

        condition = self.diabetic_data[col] == 0
        self.diabetic_data.loc[condition, col] = '?'
        self.diabetic_data['temp'] = self.diabetic_data['temp'].fillna('Other')
        condition = self.diabetic_data['temp'] == '0'
        self.diabetic_data.loc[condition, 'temp'] = np.nan
        self.diabetic_data[col] = self.diabetic_data['temp']
        self.diabetic_data.drop('temp', axis=1, inplace=True)

    self.diabetic_data.dropna(inplace=True)

    return self
```

מתודה זו אחראית להכין לקידוד את שלוש העמודות – [diag_1, diag_2, diag_3] לאחר חקירה במרשתת, מצאנו כי קודי הדיאגנוזות של עמודות אלה מייצגות סיפטומים המותאמים למחלות\בעיות ספציפיות בגוף החולה.

על פי כן, לקחנו כל ערך מאחת העמודות הללו ומיפינו אותו לתחום החולי אליו הוא מותאם, בהתאם לקוד הדיאגנוזה.

מצאנו מספר מחלות ובעיות בריאותיות המותאמות לקודים אלו, כגון: בעיות בספיקת דם, בעיות במערכת השרירים, בעיות במערכת העיכול, וסיפטומים של סכרת.

מאחר ואנו מעוניינים לחקור את תחום הסוכרתיות השארנו את הערכים בעלי הקודים המותאים לסיפטומים סוכרתיים ואת כל השאר ייצגנו כ-Other.

המתודה הבאה היינה המתודה בעלת החתימה – encode_age המוצגת להלן:

```
def _encode_age(self):
    """
    function responsible for encoding age column
    :return: class object after encoding age column
    """
    self.diabetic_data.loc[self.diabetic_data['age'] == '[0-10)', 'age'] = 5
    self.diabetic_data.loc[self.diabetic_data['age'] == '[10-20)', 'age'] = 15
    self.diabetic_data.loc[self.diabetic_data['age'] == '[20-30)', 'age'] = 25
    self.diabetic_data.loc[self.diabetic_data['age'] == '[30-40)', 'age'] = 35
    self.diabetic_data.loc[self.diabetic_data['age'] == '[40-50)', 'age'] = 45
    self.diabetic_data.loc[self.diabetic_data['age'] == '[50-60)', 'age'] = 55
    self.diabetic_data.loc[self.diabetic_data['age'] == '[60-70)', 'age'] = 65
    self.diabetic_data.loc[self.diabetic_data['age'] == '[70-80)', 'age'] = 75
    self.diabetic_data.loc[self.diabetic_data['age'] == '[80-90)', 'age'] = 85
    self.diabetic_data.loc[self.diabetic_data['age'] == '[90-100)', 'age'] = 95

    return self
```

מתודה זו אחראית לקידוד את ערכי העמודה 'age'. אין הרבה מה להסביר כאן, מאחר וערכי הגיל מיוצגים בערכים קטגורים ordinalים, מיפינו כל טווח גילאים לממוצעו.

המתודה הבאה היינה המתודה בעלת החתימה – encode_glucose המיוצגת להלן:

```
def _encode_glucose_level(self):
    """
    function responsible for encoding glucose levels in blood based on medical information
    :return: class object after encoding glucose column
    """
    max_glu_serum_dict = {'None': 0,
                          'Norm': 100,
                          '>200': 200,
                          '>300': 300
                          }
    self.diabetic_data['max_glu_serum'] = self.diabetic_data['max_glu_serum'].replace(max_glu_serum_dict)

    return self
```

מתודה זו אחראית לקודד את ערכי העמודה 'max_glu_serum'.

עמודה זו מכילה את ערכי בדיקות רמת הגלוקוז בדם של החולים אשר ביצעו בדיקה זו.

הערכים הנ"ל מייצגים את היחס בין מיליגרם גלוקוז לליטר בדם.

לאחר חקירה, מצאנו כי הערך הנורמטיבי והרצוי בדם היינו בין 70 ל130 מיליגרם גלוקוז לליטר דם. על כן, מיפינו את הערך 'norm' בעמודה זו, המייצג ערכי בדיקה נורמטיבים ל-100, הערך הממוצע של הערך הנורמטיבי של היחס בין מיליגרם גלוקוז לליטר דם.

את שאר הערכים – none, >200, >300 מפינו לכדי 0, 200, 300 בהתאמה.

המתודה הבאה היינה המתודה בעלת החתימה הבאה - encode_hemoglobin_level המוצגת להלן:

```
def _encode_hemoglobin_level(self):
    """
    function responsible for encoding HbA1c hemoglobin levels in blood based on medical information
    :return: class object after encoding HbA1c test result column
    """
    hba1c_dict = {'None': 0,
                  'Norm': 5,
                  '>7': 10,
                  '>8': 15
                  }
    self.diabetic_data['A1Cresult'] = self.diabetic_data['A1Cresult'].replace(hba1c_dict)


    return self
```

מתודה זו אחראית לקודד את ערכי העמודה 'A1Cresult'.

עמודה זו מתארת את ערכי בדיקת ההמוגלובין A1C אצל חולים אשר ביצעו בדיקה זו.

בדיקה זו מציגה את הקורולציה בין אחוזי המצאות החלבון A1C ביחס לממוצע הגלוקוז בדם.

לאחר חקירה, מצאנו כי הערכים הנורמטיבים נעים בין 4 ל5.3 ולכן נייצג את הערך 'norm' המייצג ערכי בדיקה נורמטייבים למספר העגול – 5, הערך הממוצע בעיגול בין המצאות החלבון לגובה הגלוקוז בדם.

את שאר הערכים – 8< ,7< ,0 מיפינו ל- 15 ,10 ,0 בהתאם. נתנו משקל גבוה יותר לערכים 7< ו8< מאחר ובמקרה כזה, בן האדם הנבדק חולה בסוכרת בוודאות, וככל שערך גדל החל מ-6.8 חומרת הסוכרת גדלה.

המתודה הבאה היינה המתודה בעלת החתימה – encode_medicines_change אשר מוצגת להלן :

```
def _encode_medicines_change(self):
    """
    function responsible for encoding change in medicines submission
    :return: class object after encoding medicines change column
     """
    medicines_change_dict = {'No': 0,
                             'Ch': 1
                             }
    self.diabetic_data['change'] = self.diabetic_data['change'].replace(medicines_change_dict)

    return self
```

מתודה זו אחראית לקודד את ערכי העמודה 'change'.

ערכי עמודה זו מייצגים ערכים בוליאנים - האם במהלך המצאות החולה בבית החולים היה שינוי בתרופות אשר הוא רגיל לקבל.

אין הרבה מה לפרט פה, מלבד העובדה שמיפינו את הערך 'no' אשר מייצג אי שינוי בקבלת התרופות כ-0 ואת הערך 'Ch' אשר מייצג שינוי בקבלת התרופות כ-1.

המתודה האחרונה בטיפול הידני בדאטא היינה המתודה בעלת החתימה – encode_nominal_data_to_str המוצגת להלן:

```
def _encode_nominal_data_to_str(self):
    """
    function responsible to change nominal numeric columns to string values so won't be counted at z score
    normalization, but in one-hot encoding
    :return: class object after change in those nominal numeric columns
    """
    for column in ['discharge_disposition_id', 'admission_type_id', 'admission_source_id']:
        self.diabetic_data[column] = self.diabetic_data[column].astype(str)

    return self
```

מתודה זו אחראית להכין את העמודות 'discharge_disposition_id', 'admission_type_id', 'admission_source_id' לקידוד אוטומטי.

העמודות המצוינות להלן היינן עמודות אשר מכילות ערכים מספריים.

ערכים אלו אינם ערכים מספריים בעלי מישקול סטנדרטי מספרי, אלא ערכי קוד קטגוריים nominalים- משמע ללא חשיבות לסדר המספרי אלא למשמעות הערך העומד מאחורי הקוד.

על פי כן, אנו מייצגים ערכים מספריים אלו כstringים בכדי שכאשר נבצע נורמליזציה לערכים המספריים, ערכי עמודות אלו לא ישתתפו בתהליך זה.

לאחר שהצגנו את תהליך המיפוי והקידוד הידני, נציג את תהליך הקידוד האוטומטי המוצג להלן:

```
def _automatic_encode_categorical_data(self):
    """
    automatic encoding data set with one-hot encoder and normalizing with z-score
    :return: class object after automatic encoding data
    """
    scaler = StandardScaler()
    numeric_columns = self.diabetic_data._get_numeric_data().columns
    self.diabetic_data[numeric_columns] = scaler.fit_transform(self.diabetic_data[numeric_columns])
    categorical_columns = self._get_categorical_columns()
    self.diabetic_data = pd.get_dummies(self.diabetic_data, columns=list(categorical_columns))

    return self
```

לאחר שהדאטא שלנו ממופה ומקודד ידנית ומוכן לטיפול אחרון, אנו כתחילה לוקחים את כלל העמודות המספריות ונשנה את טווח הערכים שלהם בצורה כזו שממוצע כל עמודה הינו 0 עם שונות של 1.

נגיע למטרה זו על ידי נירמול Z-Score. נירמול זה פועל בצורה הבאה: תחילה עובר על כלל הערכים בעמודה מספרית ומחשב את ה"mean" שלהם. לאחר מכן מחשב את ה"standard deviation" בעזרת ה"mean" שמצא בשלב הקודם. לאחר שחישב את פרמטרים אלו, הוא עובר על כל איבר בעמודה המספרית ומחסיר ממנו את הmean ומחלק את התוצאה בstandard deviation.

לאחר שביצענו את נירמול Z-Score, נשאר לנו לטפל בעמודות המכילות ערכים קטגוריים.

בכדי לקודד את העמודות אשר מכילות ערכים קטגוריים נפעיל את שיטת הקידוד – One-Hot. שיטה זו פועלת בצורה הבאה: לכל עמודה בעלת ערכים קטגוריים, ניקח כל ערך קטגורי uniqueי ונקצה לו עמודה משלו המקבל ערכים בוליאנים 0 ו 1. כאשר 0 מציג את העובדה כי הערך באותה רשומה לא כלל את הערך הקטגורי הנל, ו1 מייצג את ההפך. נפעל כך על כל עמודה קטגורית ונקבל עמודה ייעודית לכל ערך קטגורי.

שיטה זו (one-hot) אכן מעלה את מספר הפיצ'רים שלנו ומקשה על טיפול בדאטא אך הערכים שלנו היינם ערכים קטגוריים nominalים, מה שמוביל אותנו לשיטה זו לשם קידוד ערכים כאלו.

ניתוח ראשוני של המידע

ביצוע ניתוח ראשוני על המידע איתו אנו מתעסקים בכדי להכיר את תוכנו, ממה הוא מורכב ועל אילו פיצ׳רים נרצה לשים דגש. הסקת מסקנות ראשוניות מתבצעת על מנת שהמשך התהליך יהיה ממוקד יותר ונתחשב אך ורק בתכונות החשובות שמהם ניתן להסיק את רוב המסקנות ולהגיע לתוצאות.

כסיכום הפרק הקודם, מבחינה ראשונית עולה כאמור כי מאגר המידע מכיל מידע לא שלם וחסר, ישנן שלוש תכונות שלא נרצה להתעסק איתן ולטפל בהן עקב אחוז גבוה של ערכים חסרים – משקל (97% ערכים חסרים), קוד משלם (40% ערכים חסרים) ומומחיות רפואית (47% ערכים חסרים). תכונת המשקל דלילה מאוד ולכן לא נוכל להשתמש בה כלל, קוד המשלם הוסר מכיוון שיש בו אחוז גבוה של ערכים חסרים וחסר השפעה על התוצאות, מומחיות רפואית אומנם בעלת אחוז גבוה של ערכים חסרים אך הינה תכונה חשובה ובעלת השפעה על התוצאות ולכן לא נמחק אותה מהמאגר.

כחלק מניקוי מאגר המידע הוסרו כל המפגשים שהביאו למצב בו המטופל לא יחזור לאשפוז באופן ודאי כגון מוות של המטופל.

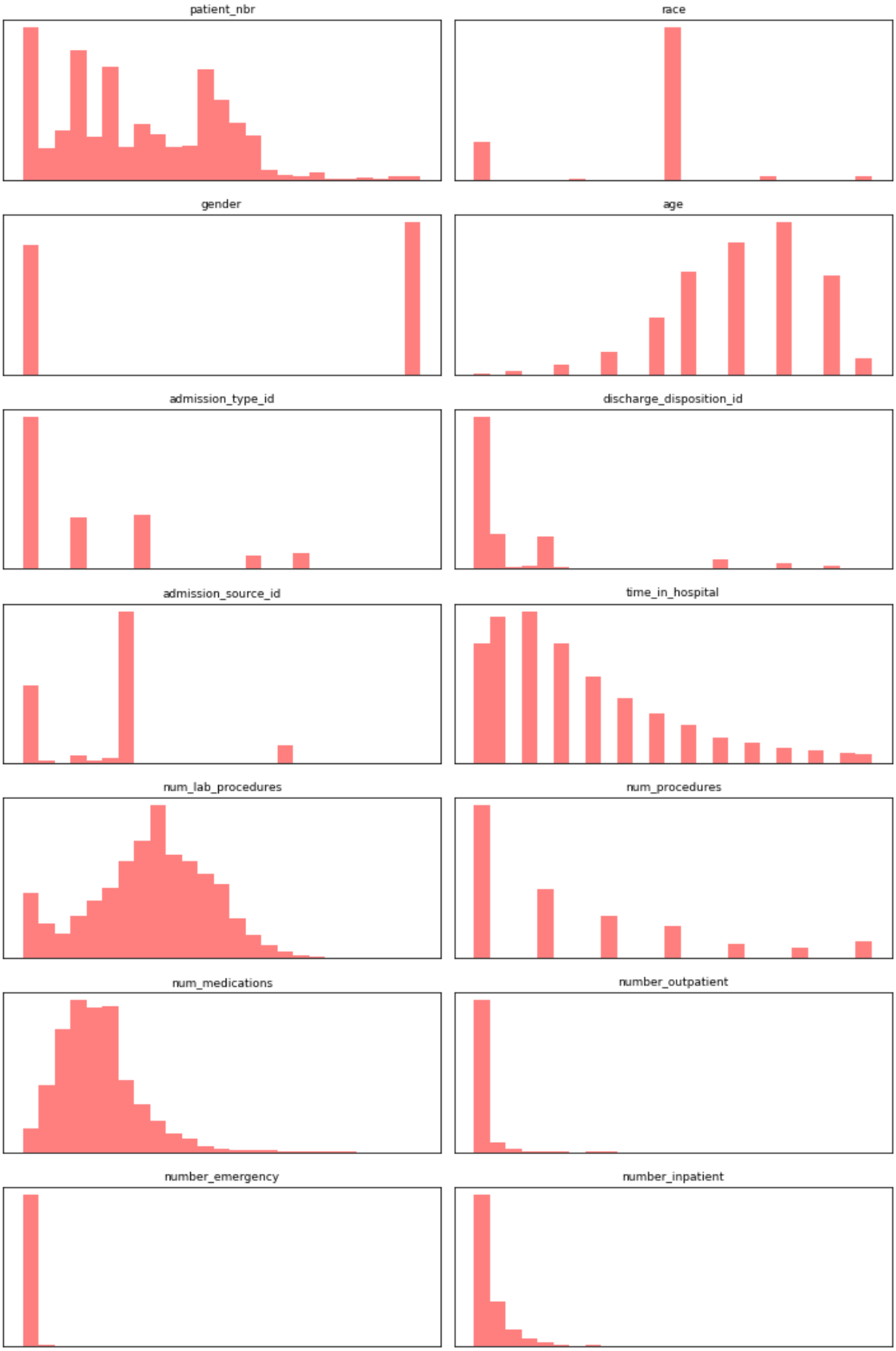
בנוסף ניתן להבין כי קיימים במאגר מטופלים בעלי פירוט עבור מספר מפגשים שלהם, דבר שיוצר תלות ופוגע במבחנים הסטטיסטיים, לכן כדי ליצור אי תלות נבחר מפגש אחד עבור כל מטופל שיהיה מוצג במאגר.

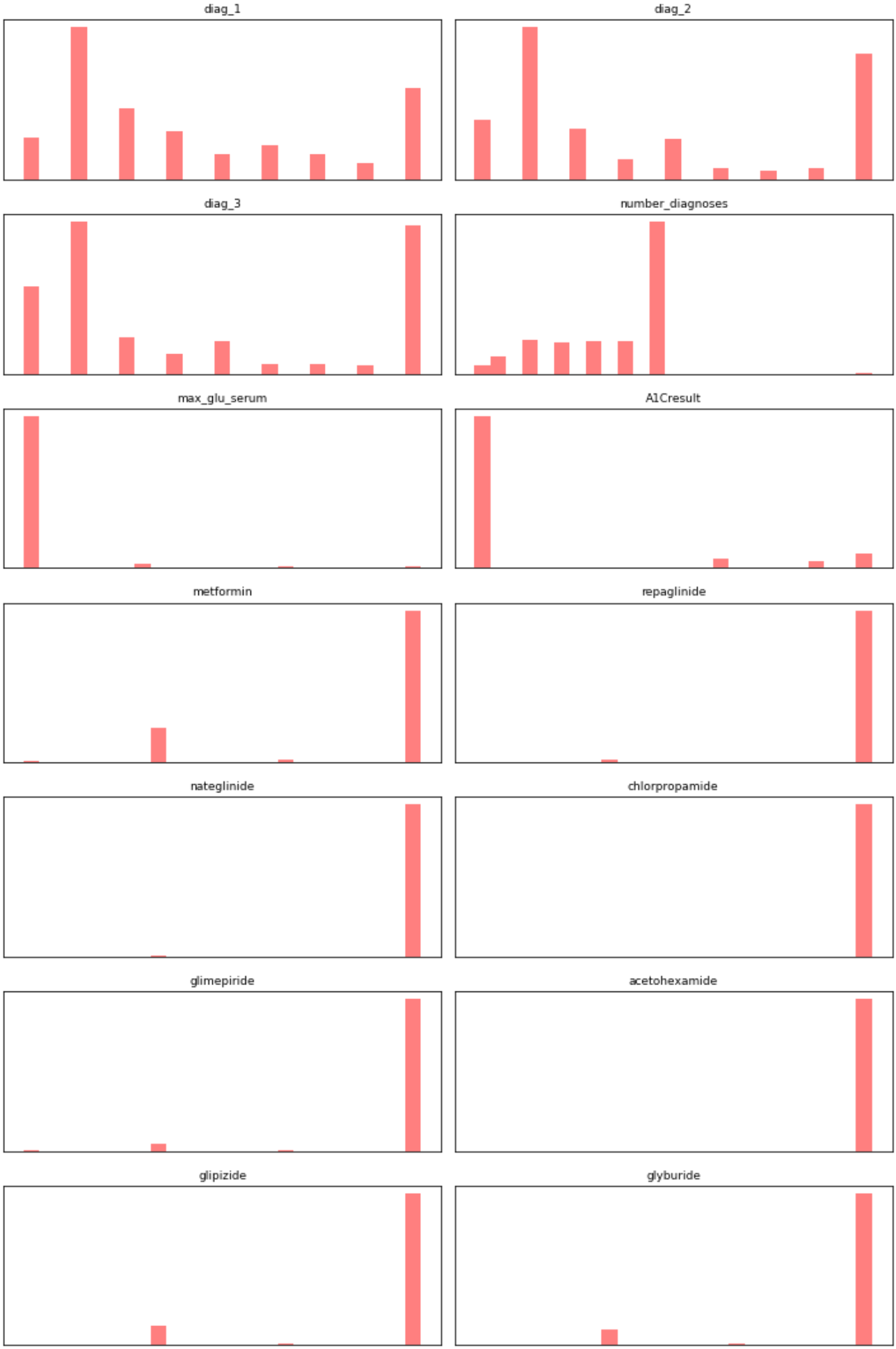
על מנת להכיר טוב יותר את מאגר המידע השתמשנו תחילה בהיסטוגרמות תוך התייחסות לשאלות המחקר שלנו.

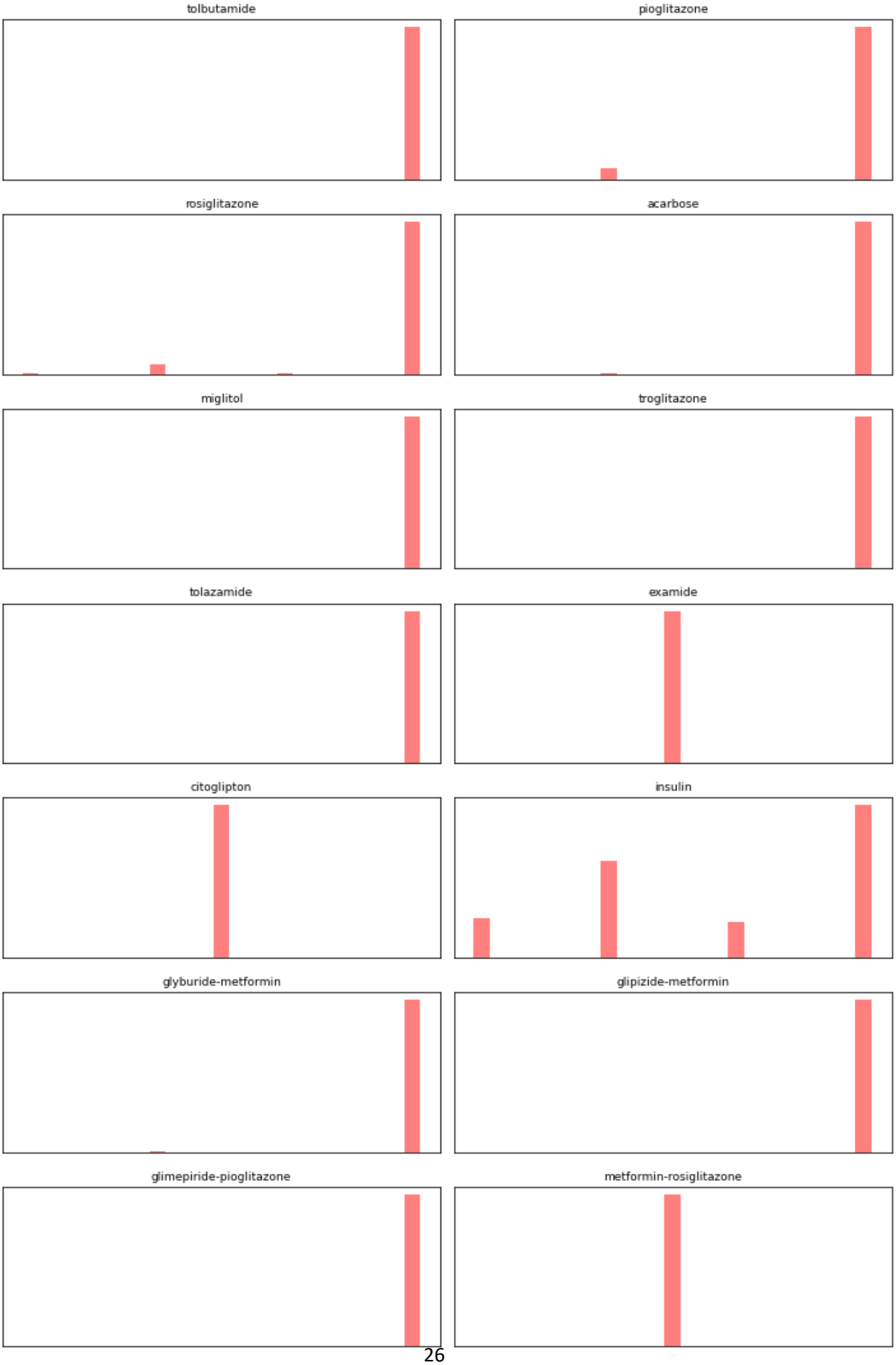
כלומר, כל היסטוגרמה מייצגת את אחד הפיצר׳ים בסט המידע שלנו – diabetic_data.csv, כאשר יש חשיבות לצבע על פי התצפיות בנוגע לאישפוזים חוזרים של המטופל בטווח זמן מסוים.

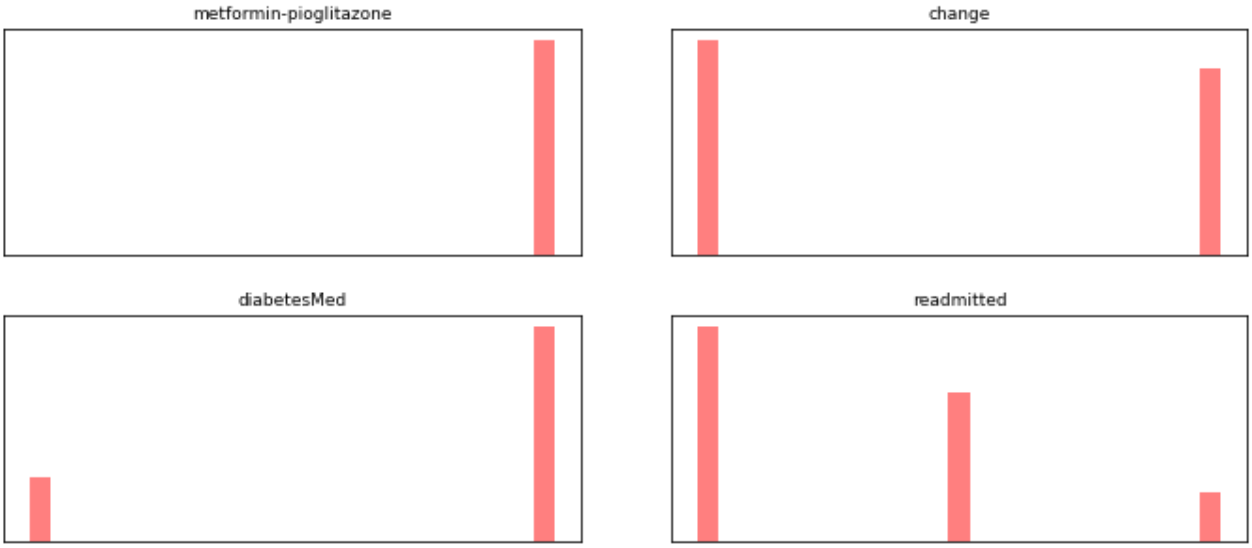
מאגר המידע מכיל פיצ׳רים בעלי ערכים מסוגים שונים (ערכים קטגוריים, ערכים מספריים ועוד) וכדי להשתמש בהיסטוגרמות בקלות המרנו את כל הפיצרים לערכים מספריים - קוד ההמרה מופיע בקובץ (research_data_processor.py).

היסטוגרמות


patient_nbr
race
gender
age
admission_type_id
discharge_disposition_id
admission_source_id
time_in_hospital
num_lab_procedures
num_procedures
num_medications
number_outpatient
number_emergency
number_inpatient


diag_1
diag_2
diag_3
number_diagnoses
max_glu_serum
A1Cresult
metformin
repaglinide
nateglinide
chlorpropamide
glimepiride
acetohexamide
glipizide
glyburide


tolbutamide
pioglitazone
rosiglitazone
acarbose
miglitol
troglitazone
tolazamide
examide
citoglipton
insulin
glyburide-metformin
glipizide-metformin
glimepiride-pioglitazone
metformin-rosiglitazone



מניתוח ראשוני של ההסטוגרמות עולה כי :

את ההיסטוגורמות שמתארות את התפלגות התצפיות בהתאם לשימוש בתרופות מסוימות ניתן לחלק לשלושה מקרים :

a. רשימת התרופות שלא היה בהם שימוש כלל/היה שימוש מועט מאוד אותו ניתן להזניח :
   i. Acetohexamide
   ii. Chlorpropamide
   iii. Tolbutamide
   iv. Miglitol
   v. acarasbose
   vi. Tolazamide
   vii. Troglitazone
   viii. Glipizide – Metformin
   ix. Metformin – Pioglitazone
   x. Nateglinide
   xi. Repaglinide
   xii. Glyburide – Metformin
   xiii. Glimepiride – Pioglitazone

ניתן להסיק כי אין שימוש בתרופות אלו בקרב חולי סוכרת, על כן אין מידע זה תורם לנו לתהליך הלימדה\ההסקה ולכן אנו יכולים להתעלם ממנו.

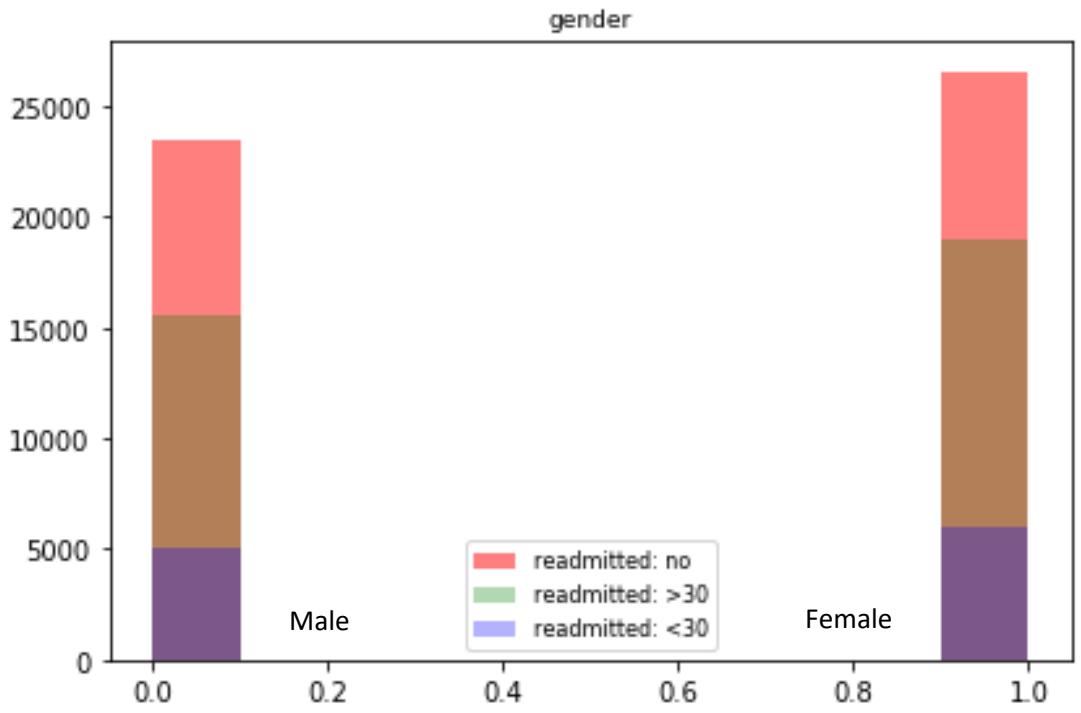
b. רשימת התרופות שבכלל התצפיות ניתן לראות שהיה בהן שימוש ללא כל שינוי:
   i. Examide
   ii. Citoglipton
   iii. Metformin – Rosiglitazone

ניתן להסיק כי ישנו שימוש גבוה בתרופות אלה בקרב חולי סוכרת, אך המינון בקרב המשתמשים נשאר קבוע ואינו משתנה ולכן שוב, אין מידע זה תורם לנו לתהליך ההסקה\הלמידה ולכן אנו יכולים להתעלם ממידע זה במהלך הניתוח.

c. בשאר התרופות ניתן לראות התפלגות מסוימת, זהו מידע אותו ניתן לנתח ולהסיק ממנו מסקנות.
נרצה לחקור תרופות אלה ולברר את צורת השפעתם על חולי הסוכרת המתועדים לנו ואיך גוררות תרופות אלו לחזרתו או לשיפור מצבו הבריאותי של בעל סממנים סוכרתיים.

לאחר ניקוי הדאטא שנמצא מיותר מהניתוח הראשוני, נרצה לחקור לעומק את התכונות המעניינות אותנו.

ניתוח מעמיק של ההיסטוגרמות

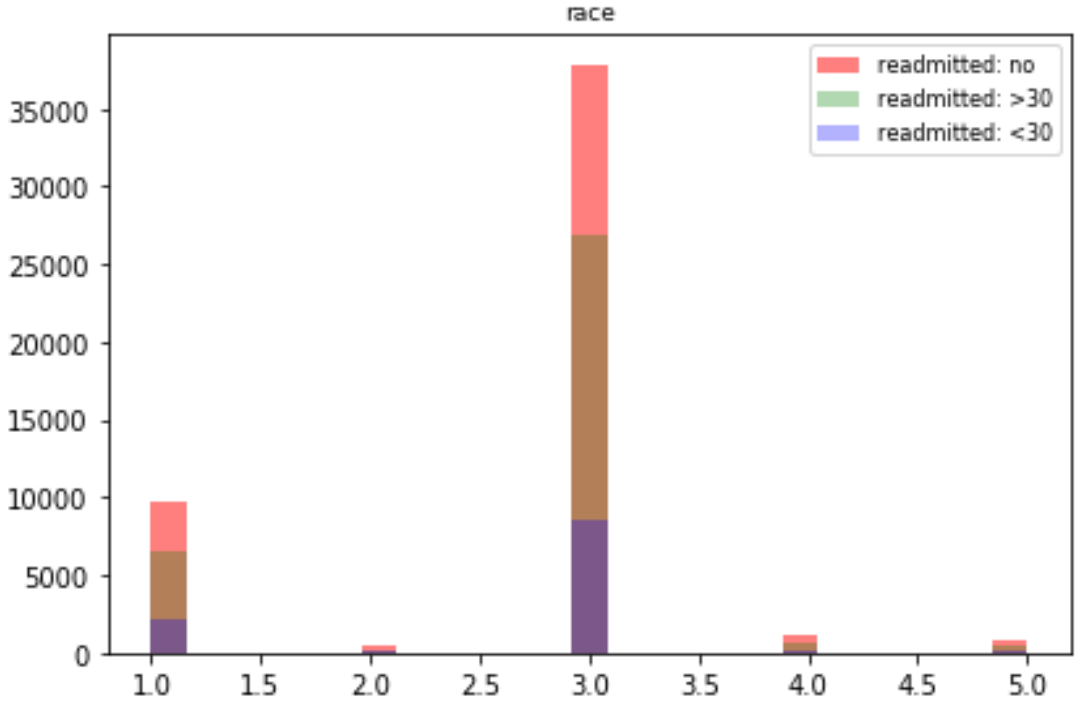


readmitted	NO	30>	30<
gender			
Male	0.468797	0.449951	0.457679
Female	0.531203	0.550049	0.542321

הטבלה הנ״ל מציגה את התפלגות אחוז הגברים לנשים כאשר אנו מבחינים בכל עמודה בנפרד תחת התיוג של העמודה.

על פי ההיסטוגרמה הנ״ל אנו רואים כי התצפיות שלנו מאוזנות אחוזית מבחינת כמות הנשים והגברים.

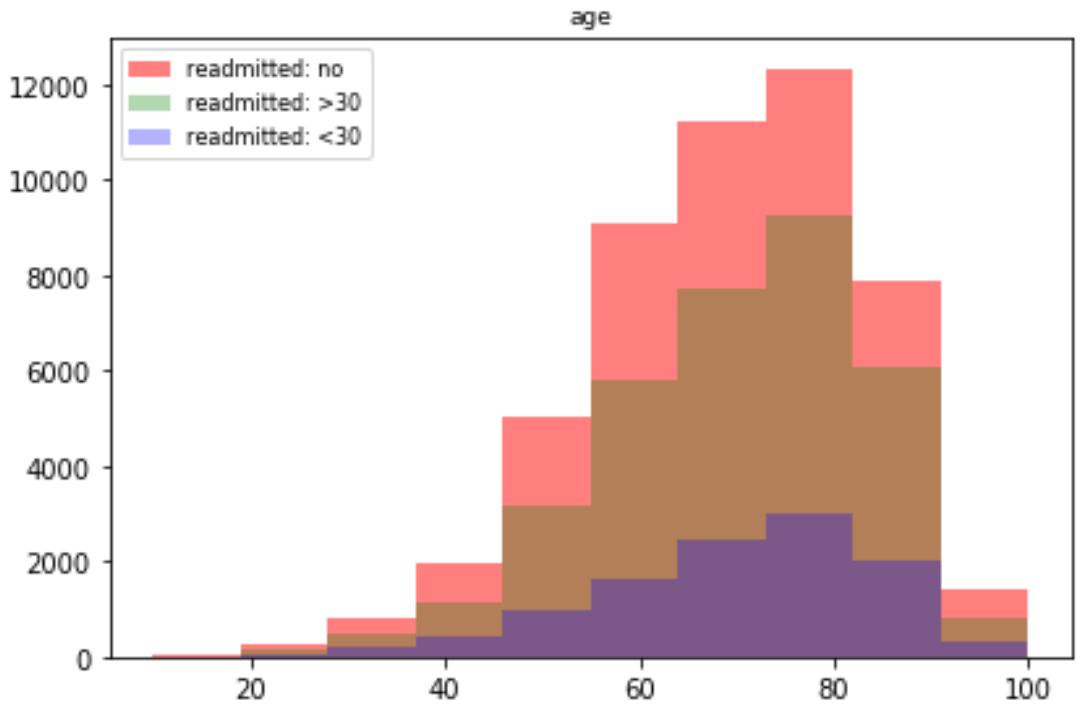
בנוסף, ניתן לראות כי אין אבחנה חד משמעית בין השפעת המין על אחת משלוש אופציות החזרה תחת העמודה 'readmitted' ואין עמודה זו משפיע על שאלת המחקר שלנו. על פי כן, אין נרצה לקחת אותה בחשבון כאשר נעבור לשלב ההסקה על ידי קלסטריזציה ולכן מחקנו אותה מן סט המידע שלנו כאמור בפרק הקודם.



other - 5	- 4 Hispanic	- 3 caucasian	- 2 Asian	- 1 AfricanAmerican	race
					readmitted
0.599034	0.572811	0.517445	0.634584	0.530416	No
0.303658	0.321792	0.366766	0.259380	0.353882	>30
0.097308	0.105397	0.115789	0.106036	0.115702	<30

>30	<30	No	readmitted
			race
0.193595	0.188609	0.195492	- 1 AfricanAmerican
0.005897	0.004595	0.007773	Asian - 2
0.768938	0.775819	0.756909	caucasian - 3
0.018779	0.018263	0.022481	Hispanic - 4
0.012791	0.012715	0.017345	other - 5

מן היסטוגרמה זו אשר מייצגת את הקשר בין הגזע של המטופל לבין חזרתו לבית החולים לאישפוז חוזר\אי חזרתו, אנו יכולים לראות שמבחינה אחוזית אנשים מגזע אסייתי מתודדים בצורה טובה ועמידה יותר למחלת הסוכרת, כמובן שאין לנו תצפיות רבות מן גזע זה היות ומדובר על תיעודים מבתי חולים מארה"ב אך קיים מידע מספיק ממנו נוכל ללמוד מסקנה זו.באם נסתכל על האדם ה"לבן" אשר מיוצג כcaucasian נראה כי עמידתו למחלת הסוכרת היינה הגרועה ביותר מכלל הגזעים.



readmitted	30>	30<	NO
age			
(0-10]	0.000264	0.000731	0.002406
(10-20]	0.003522	0.006302	0.007783
(20-30]	0.020780	0.014348	0.016605
(30-40]	0.037334	0.033394	0.039443
(40-50]	0.090429	0.092221	0.098061
(50-60]	0.146870	0.166465	0.176272
(60-70]	0.220305	0.222169	0.220254
(70-80]	0.270230	0.266564	0.246500
(80-90]	0.182971	0.175074	0.162146
(90-100]	0.027296	0.022732	0.030530

הטבלה הנ"ל מציגה את אחוז התצפיות בכל טווח גילאים בהתאם לחזרתם או אי חזרתם לבית החולים לאישפוז חוזר בטווח הזמנים המצויין בעמודה – 'readmitted'.

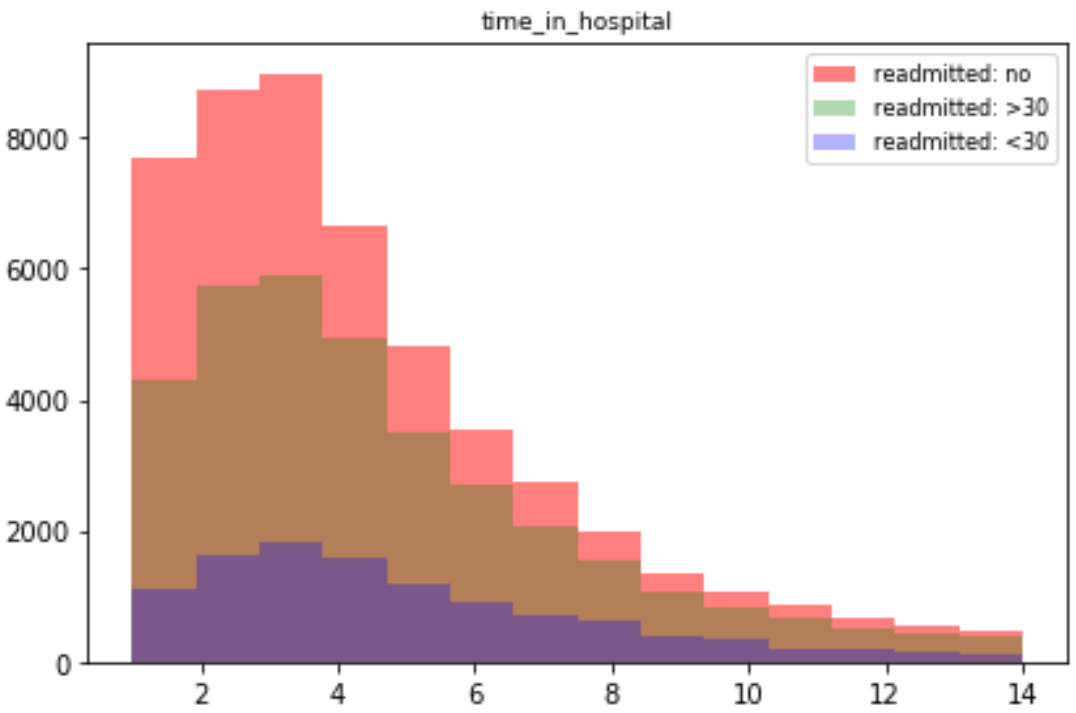
קל לראות מתוך הנתונים המספריים הנ"ל כי חולים בעלי ממצאים סוכרתיים באיזור הגילאים 60 עד 80, הם החולים אשר חוזרים בממוצע הגבוה ביותר לאישפוז חוזר בטווח הזמנים 30> (כאשר ממוצע החזרה בטווח הגילאים 60 עד 70 היינו 22 אחוז, וממוצע החזרה בטווח הגילאים 70 עד 80 היינו 27 אחוז) ביחס ל 30< (כאשר ממוצע החזרה בטווח הגילאים 60 עד 70 היינו 22 אחוז, וממוצע החזרה בטווח הגילאים 70 עד 80 היינו 27 אחוז בעיגול) מכאן נובע שאחוזי החזרה עולים ככל שמדובר באכולוסיות מבוגרות יותר. נציין כי לא התייחסנו לאופציית אי החזרה בטווח גילאים זה היות וממוצע גיל התמותה בארצות הברית לשנים אלו הוא – 73 ו78 לגברים ונשים בהתאמה.

עוד ניתן לראות כי מההיסטוגרמות הראשוניות שמציגות את הנתונים ללא חלוקה לשלוש האופציות, עולה כי התפרצות מחלת הסוכרת מתרחשת בכל הגילאים אך שכיחה בקרב אנשים מבוגרים בטווח הגילאים 60-80.

ניתן לראות ירידה חדה בכמות החולים לאחר גיל 80 וזאת ניתן להסביר על ידי ממוצע התמותה סביב ממוצע גילאים זה.

על פי ההיסטוגרמה ניתן לראות כי עמודת הגיל מתפלגת בצורה נורמלית כאשר ממוצע התצפיות הינו 71 עם סטיית תקן של 15.5. בנוסף עבור התייחסות לכל אחד מהמקרים בreadmitted בנפרד ניתן לראות כי קיימת התפלגות דומה בנוגע לגיל המטופל.

count	95672.000000
mean	71.039489
std	15.591297
min	10.000000
25%	60.000000
50%	70.000000
75%	80.000000
max	100.000000



הערכים בטבלה הנ"ל מנורמלים ע"פ readmitted

14	13	12	11	10	9	8	7	6	5	4	3	2	1	time_in_hospital
														readmitted
0.479918	0.499556	0.4805	0.4942	0.478931	0.474341	0.478	0.492233	0.494536	0.506326	0.505644	0.536816	0.542228	0.585742	NO
0.38517	0.370009	0.377483	0.396752	0.370639	0.382038	0.374609	0.376264	0.376436	0.36957	0.374195	0.353395	0.35588	0.328479	>30
0.134912	0.130435	0.142016	0.109049	0.15043	0.143621	0.147391	0.131503	0.129028	0.124104	0.120161	0.10979	0.101893	0.08578	<30

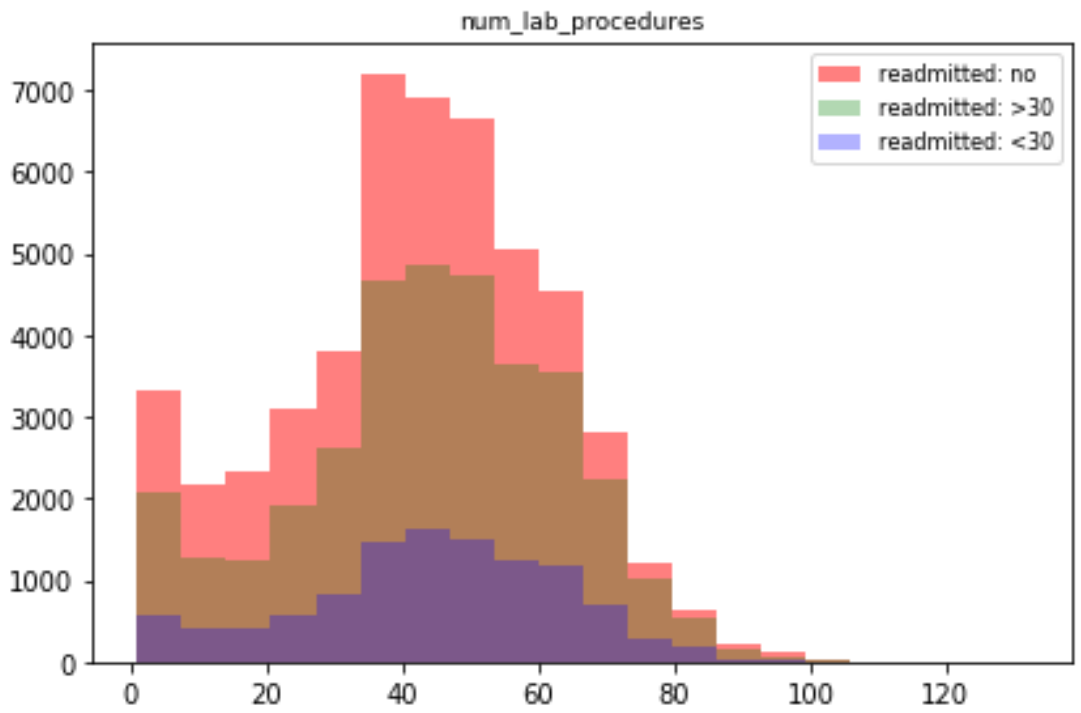
הערכים בטבלה הנ"ל מנורמלים ע"פ time in hospital

NO	30<	30>	readmitted
			time_in_hospital
0.156296	0.125784	0.102316	1
0.174851	0.166521	0.150744	2
0.178204	0.171191	0.166769	3
0.131525	0.142467	0.144757	4
0.094434	0.100886	0.105574	5
0.069900	0.077507	0.083561	6
0.054389	0.059727	0.066215	7
0.039680	0.044704	0.055032	8
0.027176	0.030919	0.036277	9
0.021289	0.023576	0.029585	10
0.017516	0.019665	0.017170	11
0.013433	0.014573	0.016994	12
0.011665	0.011844	0.013120	13
0.009642	0.010634	0.011887	14

הטבלה הנ"ל מציגה את אחוז התצפיות עבור מספר ימי האישפוז של המטופלים בבית החולים בהתאם לחזרתם או אי חזרתם לבית החולים לאישפוז חוזר בטווח הזמנים המצויין בעמודה – 'readmitted'.

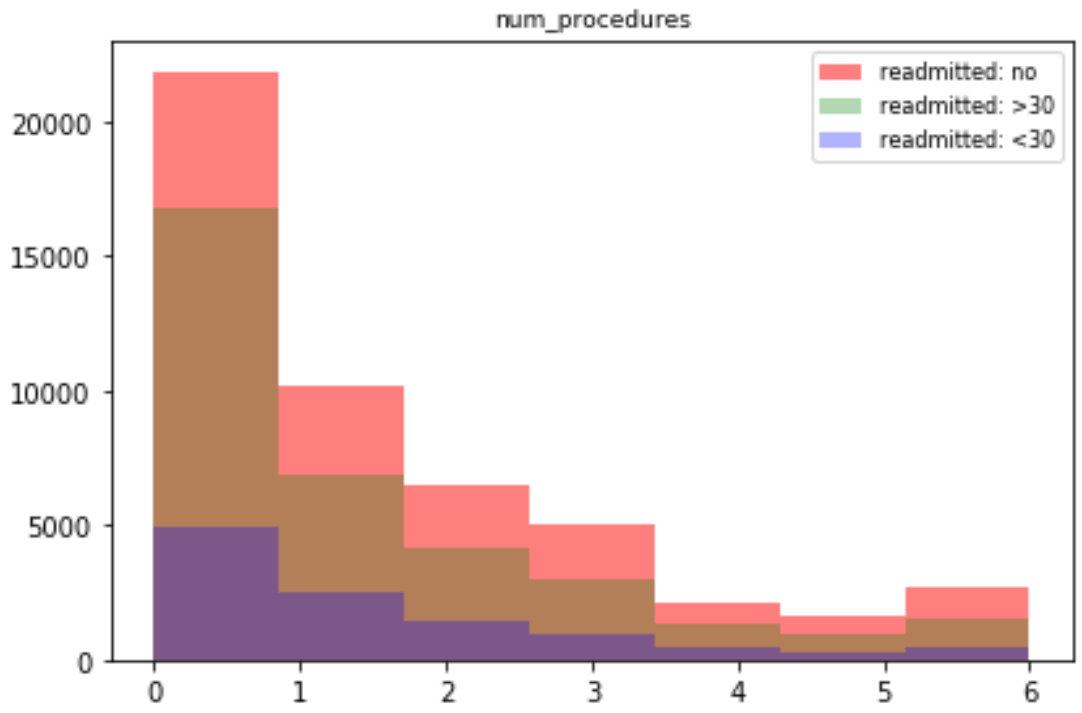
מנתונים אלה ניתן להסיק כי מחצית מהאוכלוסיה מתאשפזת לתקופה של 2-6 ימים כאשר הממוצע הינו 4 ימים עם סטיית תקן של 3 ימים. בנוסף ניתן לראות כי ככל ש מספר ימי השהייה בבית החולים גבוה יותר כך הסיכוי של המטופל לחזור לאישפוז ג דול יותר, דבר ההגיוני לוגית מאחר ואם נדרש מטופל להשאר כמות גדולה יותר של ימים באישפוז, מרמז על רמה בריאותית נמוכה הדורשת טיפול ומעקב בתוחלת זמנית גבוהה יותר.

count	95672.000000
mean	4.405155
std	2.976088
min	1.000000
25%	2.000000
50%	4.000000
75%	6.000000
max	14.000000



מנתונים אלה ניתן להסיק כי כמות בדיקות המעבדה הממוצעת עומד על 42 בדיקות עם סטיית תקן של 20 בדיקות. ניתן לראות כי מחצית מהאוכלוסיה נמצאת בתחום של 31 – 57 בדיקות.
ניתן לראות כי ישנו פיק של ערכים בבין של מספר הבדיקות בין 0 ל8, ניתן להסיק כי דבר זה קורה מפני שכאשר בן אדם נכנס לראשונה לאישפוז ישנם בדיקות בסיסיות אשר עליו לעבור.
עוד ניתן לראות כי ככלל רוב בני האדם המגיעים לאישפוז עוברים מספר בדיקות אש ר נע בין 31 ל57 בדיקות, אשר מהווים ככל הנראה את מספר הבדיקות הסטנדרטיות אשר מהוות בסיס מספק להסקת המקרה הרפואי אשר בעקבותו הגיע בן האדם לאי שפוז. אנו רואים שככל שאנו עוברים את מספר הבדיקות הסטנדרטיות מספר הרשו מות יורד, דבר המעיד על מקרים כאשר מספר הבדיקות הנ״ל לא מביאות למסקנות חד משמעיות ולכן יש צורך להמשיך ולחקור על ידי בדיקות נוספות, ניתן להסיק כי אלו מקרים של תופעות יותר מסובכות ונדירות הדורשות בדיקות פחות סטדנרטיות
.
בנוסף לכך, מניתוח סטטיסטי שעשינו על אחוזי החזרה לבית החולים לאישפוז ביח ס למספר בדיקות המעבדה קיימת מגמתיות עולה, כך שככל שמספר בדיקות המעבד ה עולה כך קיים סיכוי חזרה גבוהה יותר אל בית החולים בתוך 30 ימים החל מתאר יך השחרור.

count	95672.000000
mean	42.956246
std	19.646308
min	1.000000
25%	31.000000
50%	44.000000
75%	57.000000
max	132.000000



הערכים בטבלה הנ"ל מנורמלים ע"פ readmitted

6	5	4	3	2	1	0	num_procedures
							readmitted
0.56959	0.562305	0.540643	0.55984	0.540231	0.523047	0.501616	NO
0.330474	0.339813	0.344391	0.329285	0.344183	0.351051	0.384363	>30
0.099936	0.097883	0.114966	0.110876	0.115586	0.125902	0.11402	<30

הערכים בטבלה הנ"ל מנורמלים ע"פ num procedures

30<	30>	NO	readmitted
			num_procedures
0.451148	0.484425	0.437184	0
0.223260	0.198289	0.204304	1
0.126281	0.119777	0.130008	2
0.090266	0.085390	0.100394	3
0.041187	0.039300	0.042663	4
0.025583	0.028290	0.032372	5
0.042275	0.044530	0.053074	6

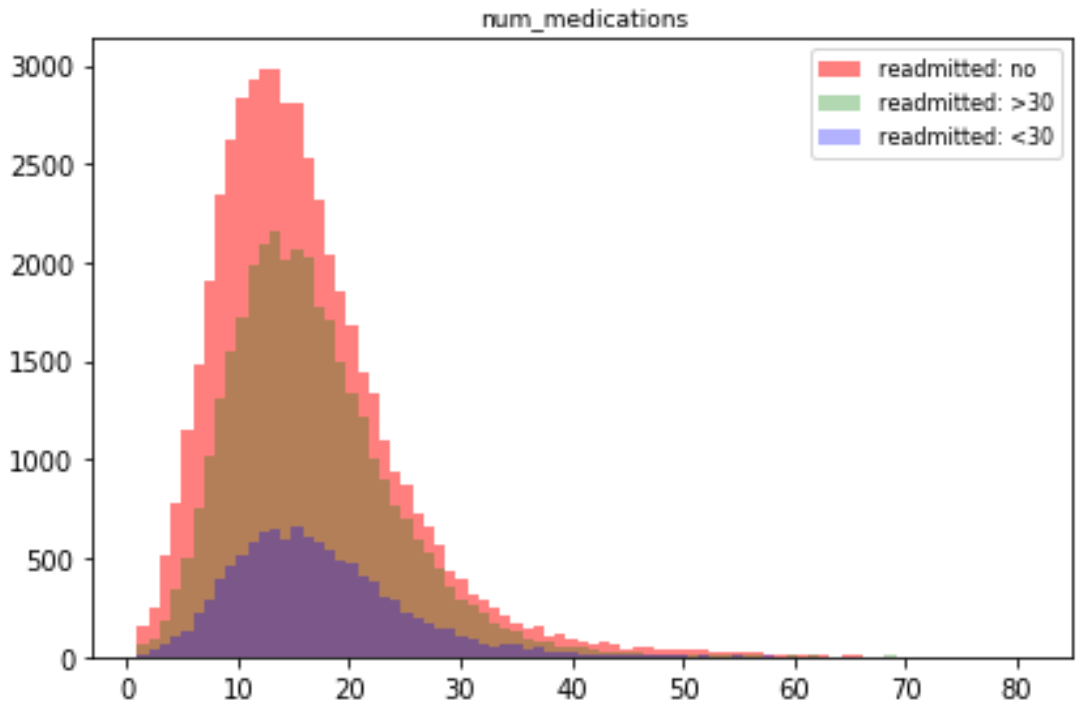
הטבלה הנ"ל מציגה את אחוז התצפיות עבור מספר הבדיקות שבוצעו בקרב המטופלים בבית החולים בהתאם לחזרתם או אי חזרתם לבית החולים לאישפוז חוזר בטווח הזמנים המצויין בעמודה – 'readmitted'.

מנתונים אלה ניתן לראות כי באופן גורף בשלוש האופציות של החזרה רוב האנשים אינם נבדקים בצורה מיוחדת. בערך NO 43% מהאנשים אינם נבדקים כלל, בערך 30> 48% מהאנשים אינם נבדקים כלל וב30< גם כן אחוז דומה של אנשים שלא אינם

נבדקים כלל 45%. בנוסף ניתן לראות דעיכה בכמות האנשים ככל שעולים במספר הבדיקות.

בנוסף מניתוח סטטיסטי של כלל נתונים אלו ללא חלוקה על פי readmitted ניתן לראות כי בממוצע מטופל עובר בדיקה אחת עם סטיית תקן של 1.7 כשאר מרבית מהמטופלים עוברים שתי בדיקות לכל היותר (75% מהמטופלים).

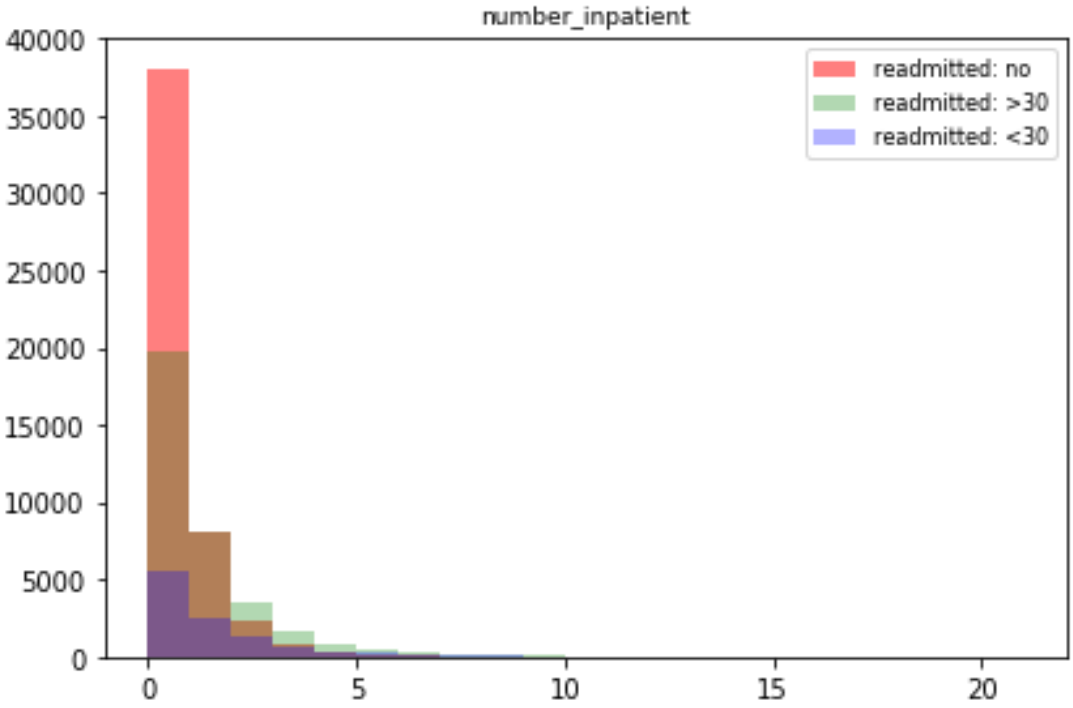
count	95672.000000
mean	1.345577
std	1.705682
min	0.000000
25%	0.000000
50%	1.000000
75%	2.000000
max	6.000000



מניתוח סטטיסטי של כלל נתונים אלו ללא חלוקה על פי readmitted ניתן לראות כי הדאטא מתפלג בצורה נורמלית עם ממוצע של 16 תרופות למטופל עם סטיית תקן של 8 תרופות. בנוסף ניתן לראות כי מרבית המטופלים מקבלים לכל היותר 20 תרופות (75% מהמטופלים).

נראה שבמקרים נדירים, קיימת נתינה של מספר רב ולא סטנדרטי של תרופות הנע בין 50 ל70 תרופות לבן אדם, דבר המעיד כנראה על מקרה מאוד מסובך (כלומר ישנו קושי במציאת תרופה אפקטיבית למקרה הנ"ל) או קשה של סוכרת.

count	95672.000000
mean	16.078278
std	8.074730
min	1.000000
25%	11.000000
50%	15.000000
75%	20.000000
max	81.000000



הערכים בטבלה הנ"ל מנורמלים ע"פ number inpatient

30<	30>	NO	readmitted
			number_inpatient
0.499592	0.568601	0.761165	0
0.223895	0.231376	0.161241	1
0.115849	0.101977	0.046940	2
0.061508	0.047680	0.017565	3
0.034020	0.023320	0.007354	4
0.022589	0.011299	0.002818	5
0.014787	0.006762	0.001419	6
0.008618	0.003468	0.000899	7
0.005987	0.001936	0.000160	8
0.004264	0.001474	0.000220	9
0.002268	0.000751	0.000140	10
0.002903	0.000462	NaN	11
0.001452	0.000376	0.000040	12
0.000907	0.000231	NaN	13
0.000363	0.000173	NaN	14
0.000726	NaN	NaN	15
0.000091	0.000058	0.000040	16
NaN	0.000029	NaN	18
0.000091	0.000029	NaN	19
0.000091	NaN	NaN	21

הערכים בטבלה הנ"ל מנורמלים ע"פ readmitted

8	7	6	5	4	3	2	1	0	number_inpatient
									readmitted
0.056738	0.173077	0.151709	0.180538	0.237419	0.274088	0.328302	0.435127	0.601991	NO
0.475177	0.461538	0.5	0.50064	0.520645	0.5145	0.493222	0.431784	0.310976	<30
0.468085	0.365385	0.348291	0.318822	0.241935	0.211413	0.178477	0.133089	0.087033	>30
	16	15	14	13	12	11	10	9	
	0.4	NaN	NaN	NaN	0.064516	NaN	0.12069	0.100917	NO
	0.4	NaN	0.6	0.444444	0.419355	0.333333	0.448276	0.46789	<30
	0.2	1	0.4	0.555556	0.516129	0.666667	0.431034	0.431193	>30

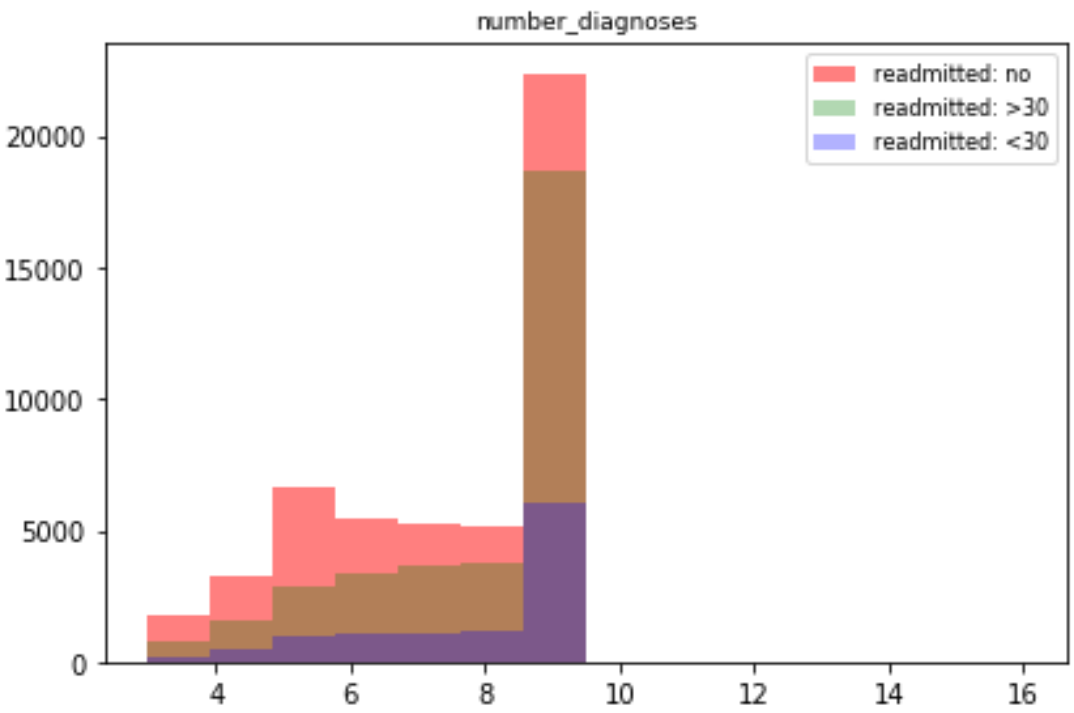
הטבלה הנ"ל מציגה את אחוז התצפיות עבור מספר אישפוזים דחופים של המטופלים בבית החולים בהתאם לחזרתם או אי חזרתם לבית החולים לאישפוז חוזר בטווח הזמנים המצויין בעמודה – 'readmitted'.

מנתונים אלו ניתן להסיק כי ככל שמטופל חזר לאישפוז בבית החולים מספר רב יותר של פעמים כך גדל הסיכוי של אותו מטופל לחזור לאישפוז בטווח זמן קצר.

מניתוח סטטיסטי של כלל נתונים אלו ללא חלוקה על פי readmitted ניתן לראות כי כמות הפעמים שמטופל מגיע לאשפוז בממוצע הינו 1, עם סטיית תקן של 1.2, כאשר 75% ממטופלים חוזרים לכל היותר פעם אחת לאישפוז בבית החולים בעיקבות אירוע.

בנוסף, אנו רואים מן המידע הסטטיסטי כי מספר האנשים שחוזרים לבית החולים יותר מפעם אחת במסגרת של אישפוז דחוף היינו נמוך מאוד. דבר המעיד על כך שהידע סביב מחלת הסוכרת וצורת הטיפול בה היינו גבוהה יחסית, בצורה כזו שסיכוי חזרת החולים לבית החולים במסגרת אירוע חריג שואף לאפס.

count	95672.000000
mean	0.642278
std	1.268598
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	21.000000



הערכים בטבלה הנ"ל מנורמלים ע"פ number diagnoses

30<	30>	NO	readmitted
			number_diagnoses
0.018325	0.021730	0.035689	3
0.040279	0.045426	0.066063	4
0.087816	0.083251	0.132966	5
0.094439	0.096660	0.108507	6
0.099519	0.105762	0.104850	7
0.110587	0.107785	0.103671	8
0.547582	0.538346	0.447255	9
0.000272	0.000116	0.000160	10
0.000272	0.000144	0.000060	11
0.000091	0.000087	0.000080	12
0.000272	0.000173	0.000140	13
0.000091	0.000058	0.000060	14
0.000091	0.000087	0.000080	15
0.000363	0.000376	0.000420	16

הערכים בטבלה הנ"ל מנורמלים ע"פ readmitted

9	8	7	6	5	4	3	number_diagnoses
							readmitted
0.475727	0.511788	0.52449	0.553178	0.633533	0.621195	0.651825	NO
0.395979	0.367959	0.365854	0.34077	0.274303	0.295378	0.274453	>30
0.128295	0.120253	0.109656	0.106051	0.092164	0.083427	0.073723	<30
16	15	14	13	12	11	10	
0.552632	0.5	0.5	0.4375	0.5	0.272727	0.533333	NO
0.342105	0.375	0.333333	0.375	0.375	0.454545	0.266667	>30
0.105263	0.125	0.166667	0.1875	0.125	0.272727	0.2	<30

הטבלה הנ"ל מציגה את אחוז התצפיות עבור מספר האבחנות שאובחנו אצל המטופלים בהתאם לחזרתם או אי חזרתם לבית החולים לאישפוז חוזר בטווח הזמנים המצויין בעמודה – 'readmitted'.

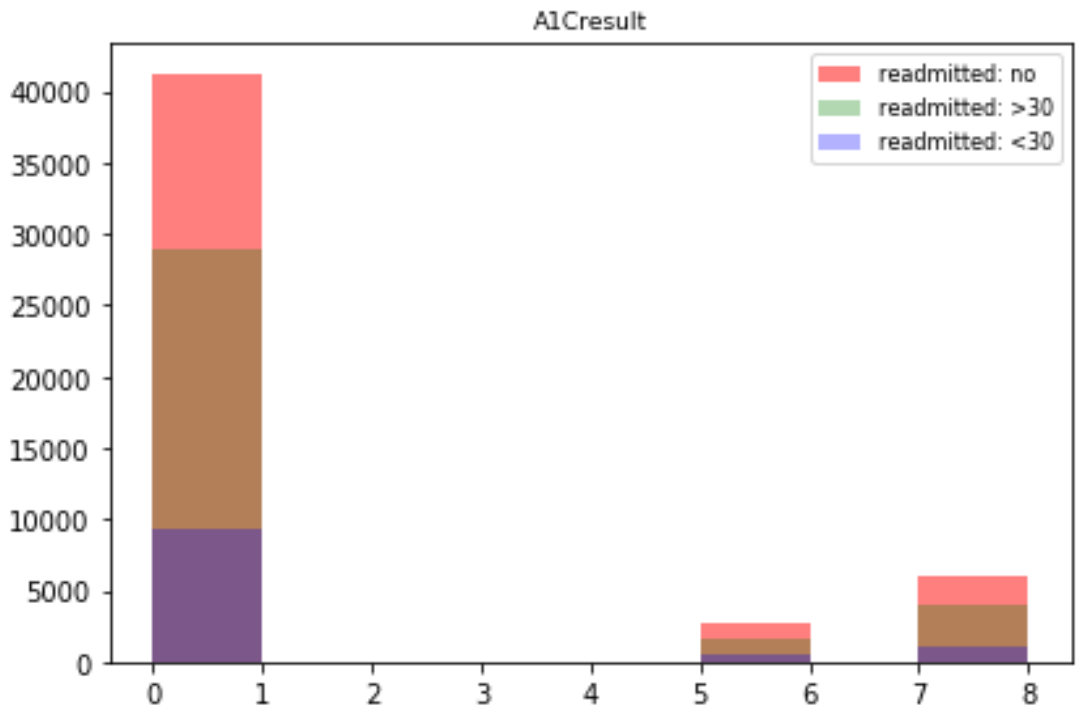
מנתונים אלו ניתן להסיק כי כמחצית מהמטופלים בכל אחת משלוש אופציות החזרה, אובחנו ב9 אבחנות נפרדות. ניתן לראות כי כמות זו מהווה חסם עליון.

מספר האנשים בעלי כמות אבחנות הגדולה מ9 מזערית מאוד המהווה פחות מאחוז אחד 0.01% אשר מהווים דגימות חריגות. בנוסף, ניתן להסיק כי ככל שכמות האבחנות של המטופלים גדולה יותר כך גדל הסיכוי שלהם לחזור לבית החולים בפרק זמן קצר; זאת ניתן לראות על פי הטבלה השנייה, על פיה אחוז האנשים שחוזרים בפרק זמן של 30 ימים גדל ככל שכמות האבחנות גדולה יותר.

מניתוח סטטיסטי של כלל נתונים אלו ללא חלוקה על פי readmitted ניתן לראות כי עם ממוצע התצפיות עבור מספר אבחנות הינו 7 עם סטיית תקן של 2. ניתן לראות כי 75% מהאוכלוסיה בעלי 9 אבחנות לכל היותר.

ניתן לראות כי כלל המטופלים מאובחנים ב9 אבחנות, דבר המעיד על גורמים משותפים המושפעים על ידי סיפטומי הסוכרת.

count	95672.000000
mean	7.492913
std	1.839197
min	3.000000
25%	6.000000
50%	8.000000
75%	9.000000
max	16.000000



הערכים בטבלה הנ"ל מנורמלים ע"פ readmitted

A1Cresult	0	5	7	8
readmitted				
NO	0.518271	0.574593	0.550531	0.527667
30>	0.363779	0.326578	0.348406	0.368679
30<	0.117950	0.098830	0.101062	0.103654

הערכים בטבלה הנ"ל מנורמלים ע"פ A1Cresult

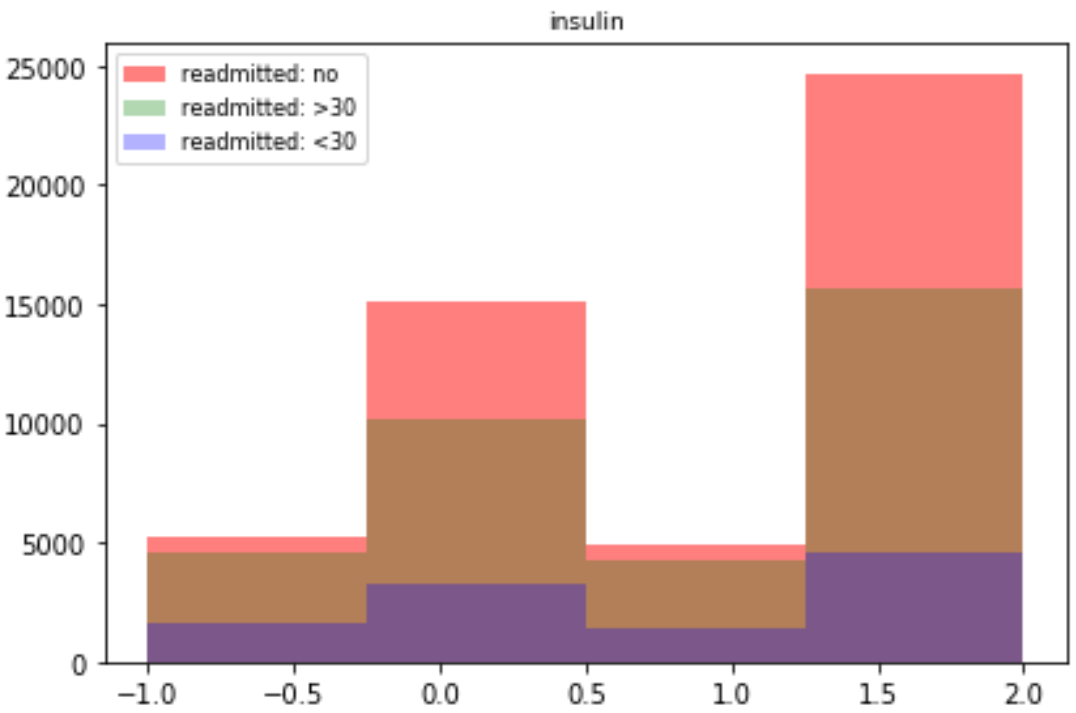
readmitted	NO	30>	30<
A1Cresult			
NO - 0	0.825010	0.837398	0.852400
Normal - 5	0.054953	0.045166	0.042910
7%> – 7	0.040385	0.036959	0.033657
8%< - 8	0.079651	0.080477	0.071033

הטבלה הנ"ל מציגה את אחוז התצפיות עבור תוצאות הבדיקה A1C שמודדת את מידת ההמוגלובין, בהתאם לחזרתם או אי חזרתם לבית החולים לאישפוז חוזר בטווח הזמנים המצויין בעמודה – 'readmitted'.

על פי ניתוח סטטיסיטי שנעשה על הדאטא ניתן לראות כי רק 16.7% מהמטופלים עשו את הבדיקה לגילוי רמת הסוכר (גלוקוז) הממוצעת בדם במשך שלושת החודשים האחרונים. בנוסף ניתן לראות כי 47% מתוך המטופלים שעברו את הבדיקה נמצאו בעלי אחוז גבוה של גלוקוז בדם (יותר מ8%), כלומר כמעט מחצית מהאנשים שעשו את הבדיקה נמצאו בעלי אחוז גבוה בהרבה מהנורמה.

נראה כי רוב מן המטופלים אינו עובר בדיקה זו, הסיבה מאחורי תופעה זו היינה העובדה כי בדיקה זו היינה בדיקת סף לווידוא מחלת הסוכרת. לכן נסיק כי רוב המטופלים המגיעים לאישפוז היינם מטופלים אשר כבר מאובחנים בחיוביות במחלת הסוכרת, כאשר מעטים מן המגיעים היינם מאובחנים חדשים. יש לציין שלא מדובר על מאובחנים חדשים בלבד, מאחר ובדיקה זו מומלצת פעם בשנה גם לאנשים המאובחנים כבר בסוכרת.

count	95672.000000
mean	1.150378
std	2.619305
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	8.000000



הערכים של הבינים בהיסטוגרמה מייצגים את מינון התרופה עבור המטופלים.
הערך 1- מייצג הורדה במינון התרופה, 0 – מינון התרופה נשאר קבוע, 1 – מינון התרופה עלה ו2 – אי שימוש בתרופה על ידי המטופל.

הערכים בטבלה הנ"ל מנורמלים ע"פ readmitted

insulin	Down	Steady	Up	No
readmitted				
NO	0.455853	0.530019	0.466433	0.549226
30>	0.400659	0.354593	0.398449	0.347607
30<	0.143488	0.115389	0.135117	0.103168

הערכים בטבלה הנ"ל מנורמלים ע"פ insulin

readmitted	NO	>30	<30
insulin			
Down	0.105130	0.133618	0.150231
Steady	0.302900	0.293042	0.299374
Up	0.098575	0.121771	0.129638
No	0.493396	0.451569	0.420757

הטבלה הנ"ל מציגה את אחוז התצפיות עבור צריכת אינסולין, בהתאם לחזרתם או אי חזרתם לבית החולים לאישפוז חוזר בטווח הזמנים המצויין בעמודה – 'readmitted'.

מהנתונים הנ"ל ניתן להסיק כי שינוי במינון התרופה גרם לסיכוי גבוה יותר של חזרה לבית החולים בפרק זמן קצר בקרב מטופלים שצורכים תרופה זו.

מניתוח סטטיסטי עולה כי 45% אחוז מהמטופלים לא צורכים תרופה זו. בנוסף ל56% מהמטופלים שמשתמשים בתרופה זו לא השתנה המינון, ל22% מהם המינון ירד ולשאר הוא עלה 20.8%.

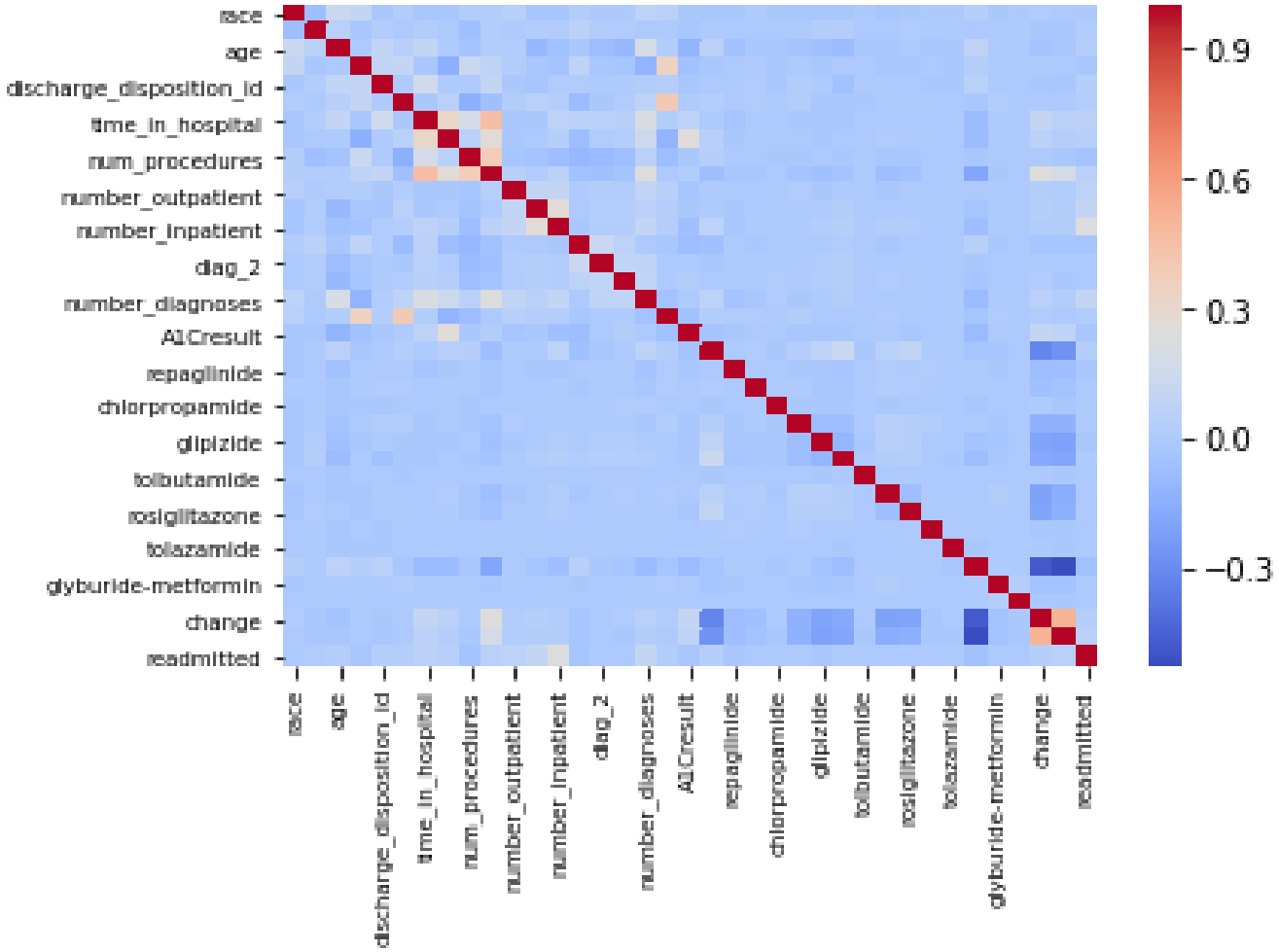
מציאת מתאמים (קורלציות)

מציאת קורלציות וקשרים בין כל שני פיצרים נועדו כדי לתת לנו אינדיקציה לתכונות חשובות ומעניינות אותן נרצה לחקור ולנתח.

על פי טבלת ציוני הקורלציה (מצורף בקובץ correlation.csv תחת תקיית resources), ניתן לראות כי ניתן למצוא רק מתאמים חיובים חלשים ובינונים כלומר הציונים הגבוהים ביותר בטבלת הקורלציות של הדאטא שלנט נעים בטווח 0.2 – 0.6, מכאן ניתן להבין כי אין קשרים מובהקים וברורים בין הפיצרים אך ניתן לנתח את הדאטא ולהסיק מסקנות מסוימות.

טבלה זו מציגה את מידת הקורלציה בין כל שני פיצ'רים. הצבע נותן לנו אינדיקציה האם קיימת קורלציה בין שני פיצ'רים נתונים ומה מידתה וזאת על פי מידת החום של הצבע, כלומר צבע אדום

משמעו קורלציה מושלמת וככל שמתקרבים לצבעים קרים יותר כך הקורלציה קטנה.



בפרק הנ״ל נציין בפירוט את הקורלציות בין כל שתי עמודות העומדות בתנאים שצוינו למעלה, כלומר ציון הקורלציה בינהם נע בין 0.2 – 0.6 ונרצה לנתח אותם לעומק ולמצוא את הקשר בינהם.

נציין כי ערכי הקורולציה המוצגים הינם ערכים לאחר ביצוע נירמול עמודתי.

Insulin – readmitted

עבור העמודה בעלת ציון הקורלציה הגבוה ביותר 0.52- אנו רואים כי אין קשר מובהק ולא ניתן להסיק דבר.

diabetesMed	No	Yes
insulin		
Down	NaN	0.156522
Steady	NaN	0.387867
Up	NaN	0.143435
No	1.0	0.312176

Change – readmitted

כך גם עבור שני פיצ׳רים אלו בעלי ציון גבוהה יחסית של 0.50 אנו רואים כי אין קשר מובהק ולא ניתן להסיק דבר.

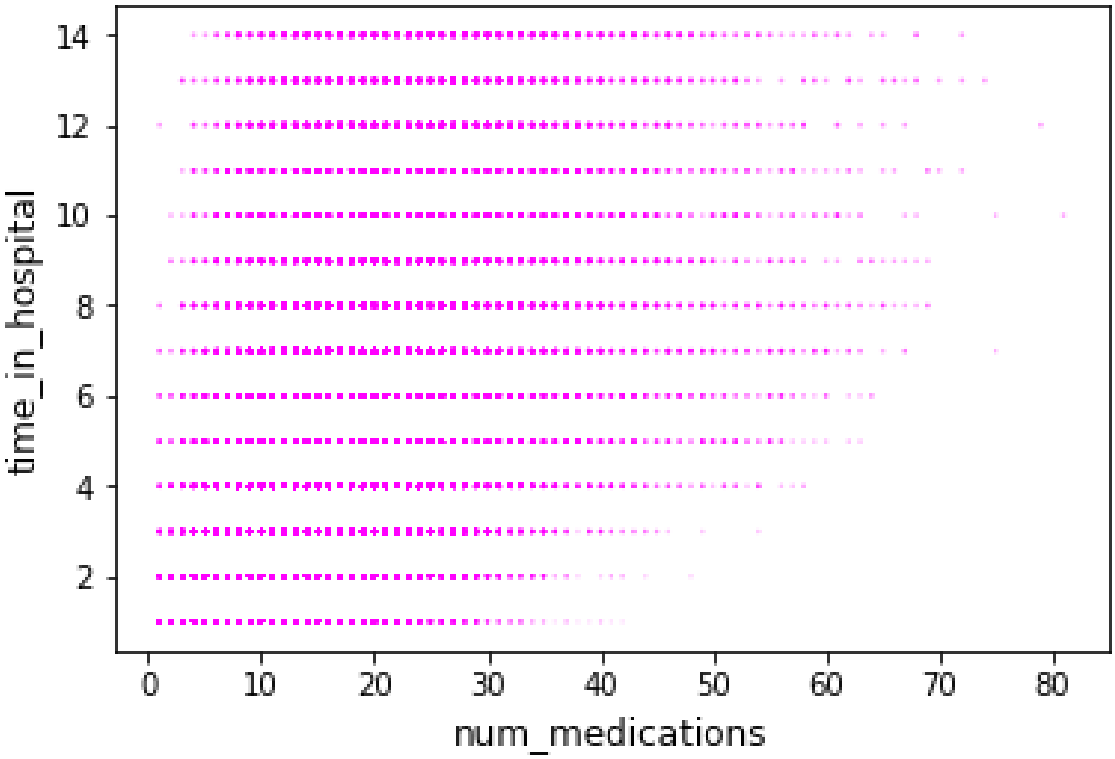
diabetesMed	No	Yes
change		
No	1.0	0.397754
Yes	NaN	0.602246

Insulin – change

ניתן לראות על פי טבלה זו כי כמחצית מהמטופלים לא שינו את הרגלי צריכת התרופה – אינסולין, בנוסף ניתן לראות כי אחוז המטופלים שהעלו את מינון התרופה ואחוז המטופלים שהורידו את מינון התרופה כמעט זהה. מנתונים אלו ניתן ללמוד על פיזור האוכלוסיה ביחס לצריכת התרופה אך לא ניתן ללמוד על מידת ההשפעה של שתי התכונות.

change	No	Yes
insulin		
Down	NaN	0.259897
Steady	0.289744	0.309530
Up	NaN	0.238166
No	0.710256	0.192406

Num Medications - time in hospital



טבלה זו מציגה את אחוז המטופלים שנמצאים בכל אחת מהאפשריות, כאשר ככל שהצבע אדום יותר כך כמות התצפיות בעלות טווח נתונים קרוב הקובע את תלות בין שהייה בבית חולים לכמות התרופות שצורך המטופל גבוהה יותר.

80	70	60	50	40	30	20	10	num_medications time_in_hospital
NaN	NaN	NaN	NaN 2.19E-10	4.81E-09	7.80E-08	6.36E-07	7.14E-07	1
NaN	NaN	NaN	NaN 6.56E-10	8.74E-09	1.40E-07	9.36E-07	6.75E-07	2
NaN	NaN	NaN 1.09E-10	1.86E-09	2.16E-08	2.45E-07	1.06E-06	4.98E-07	3
NaN	NaN	NaN 1.97E-09	6.01E-09	3.54E-08	2.71E-07	8.44E-07	2.84E-07	4
NaN	NaN 2.19E-10	4.26E-09	9.61E-09	3.65E-08	2.43E-07	5.81E-07	1.62E-07	5
NaN	NaN 4.37E-10	3.71E-09	1.05E-08	3.97E-08	2.14E-07	4.14E-07	9.81E-08	6
NaN 1.09E-10	9.83E-10	5.13E-09	1.02E-08	4.42E-08	1.85E-07	3.00E-07	5.98E-08	7
NaN	NaN 1.75E-09	4.59E-09	1.09E-08	4.35E-08	1.49E-07	2.04E-07	4.02E-08	8
NaN	NaN 1.09E-09	2.29E-09	8.96E-09	3.31E-08	1.10E-07	1.30E-07	2.03E-08	9
1.09E-10	1.09E-09	3.71E-09	9.40E-09	3.18E-08	8.88E-08	8.96E-08	1.65E-08	10
NaN 1.09E-10	9.83E-10	2.51E-09	7.43E-09	2.98E-08	6.64E-08	6.93E-08	1.18E-08	11
NaN 1.09E-10	5.46E-10	2.19E-09	5.46E-09	2.39E-08	5.53E-08	5.26E-08	8.41E-09	12
NaN 2.19E-10	1.31E-09	1.97E-09	5.90E-09	2.21E-08	4.35E-08	4.16E-08	6.56E-09	13
NaN 1.09E-10	7.65E-10	1.97E-09	6.88E-09	1.97E-08	3.92E-08	3.29E-08	4.59E-09	14

גרף זה מציג את כמות התרופות שמטופל צורך כפונקציה של מספר ימי אישפוז בבית החולים, כאשר באיזורים בהם הנקודות מוצגות בצורה עבה יותר, כמות התצפיות גדולה יותר.

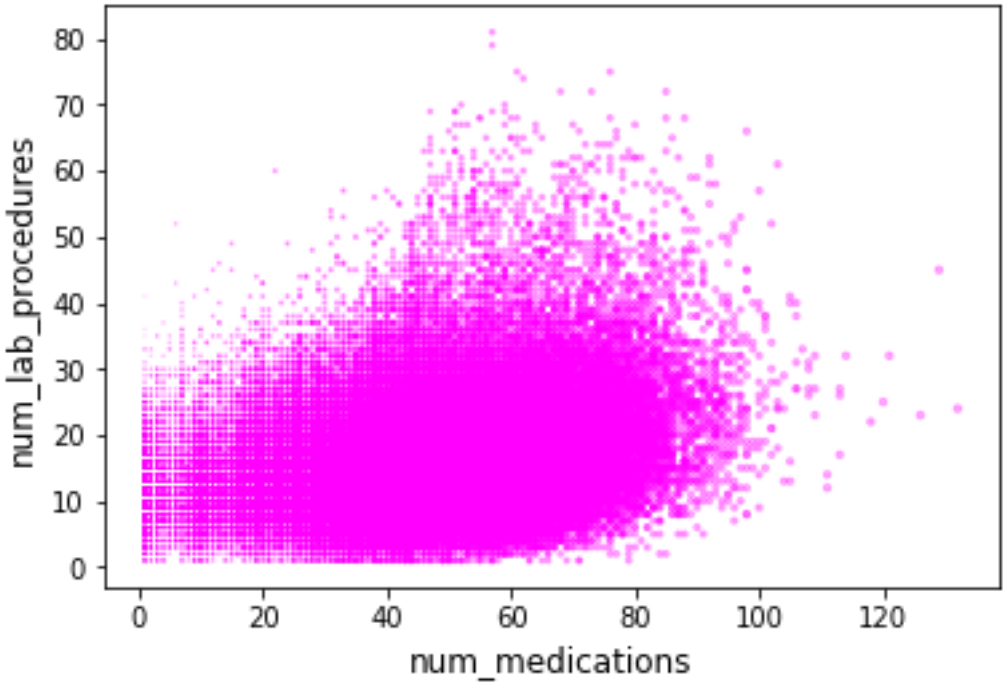
מן המידע הנ״ל נסיק כי ככל שמטופל צורך יותר תרופות, כך שהייתו בבית החולים מתארכת, ניתן לראות כי כמות המטופלים שצורכים כמות גדולה של תרופות קטנה יחסית, אך כלל מטופלים אלה מאושפזים זמן רב בבית החולים. לדוגמא, עבור מטופלים שצורכים בין 80-70 תרופות, ניתן לראות כי כמות הזמן המינימלית שהמטופל שוהה בבית החולים עומדת על 7 ימים.

מעבר לכך, ניתן לראות כי עבור צריכת כמות קטנה של תרופות כ- 5 תרופות ופחות, המטופלים מתאשפזים בבית החולים לתקופה קצרה יחסית שלא עולה על 7 ימי אישפוז.

עוד נשים לב כי חתך אוכלוסיות המטופלים מתרכז מסביב לערכי ה40-0 תרופות ומידת הפיזור בתחומים אלה עבור מספר ימי אישפוז בבית החולים הוא גדול ומתפרס על כל התחום במידה שווה.

ניתן לראות על פי מיפוי בנתונים בעזרת טבלת חום, כי כאשר נתייחס לעמודת זמן בבית החולים נראה כי רובו המוחלט של אכולוסיית המטופלים נעה בין הישארות של יום עד שישה ימים בבית החולים באישפוז, כאשר הדומיננטיות נעה בין יום ליומיים. כאשר נתייחס לעמודת מספר התרופות נראה כי צריכת התרופות בקר המטופלים נעה בין 0 ל30 תרופות, כאשר הצריכה הדומיננטית נעה בין 0 ל20 תרופות למטופל. באם נסתכל על הקורולציה בין שתי עמודות אלו נוכל להבחין כי ישנה קורולציה דומיננטית המייצגת כי מטופלים אשר נשארו באישפוז עד יומיים צרכו בממוצע עד 20 תרופות.

Num medications – num lab procedures



כל תא בטבלה הנ״ל מצייג את קבוצת האנשים שצרכו כמות מסוימת של תרופות בהתאם לכמות בדיקות המעבדה שעברו.

חילקנו את הטבלה לשני חלקים, המיוצגים על ידי שני סוגי הגוונים המופיעים (אדום וכתום). כאשר ככל שהצבע כהה יותר כך כמות האנשים גדולה יותר. החלוקה נועדה על מנת להפריד את הדאטא הנ״ל לשתי קבוצות: קבוצה המכילה את רוב המטופלים המוסמנים בצבע אדום והקבוצה השנייה, החריגה יותר בה ניתן לראות כי כמות האנשים קטנה מאוד בצבע כתום.

81	80	70	60	50	40	30	20	10	num_medications
									num_lab_procedures
NaN	NaN	NaN	1	3	44	681	4070	3189	10
NaN	NaN	NaN	NaN	10	93	768	3062	1861	20
NaN	NaN	NaN	1	16	198	1460	4843	2611	30
NaN	NaN	NaN	10	53	476	2639	8739	5116	40
NaN	NaN	9	55	193	761	3790	11101	5112	50
1	1	29	106	180	662	3343	8777	3540	60
NaN	3	15	73	169	642	2808	5932	1831	70
NaN	2	21	40	129	466	1526	2212	448	80
NaN	1	6	21	87	198	481	508	76	90
NaN	NaN	3	7	18	60	129	90	12	100
NaN	NaN	1	1	1	12	20	7	NaN	110
NaN	NaN	NaN	NaN	NaN	1	4	3	NaN	120
NaN	NaN	NaN	NaN	1	1	1	NaN	NaN	130
NaN	NaN	NaN	NaN	NaN	NaN	1	NaN	NaN	132

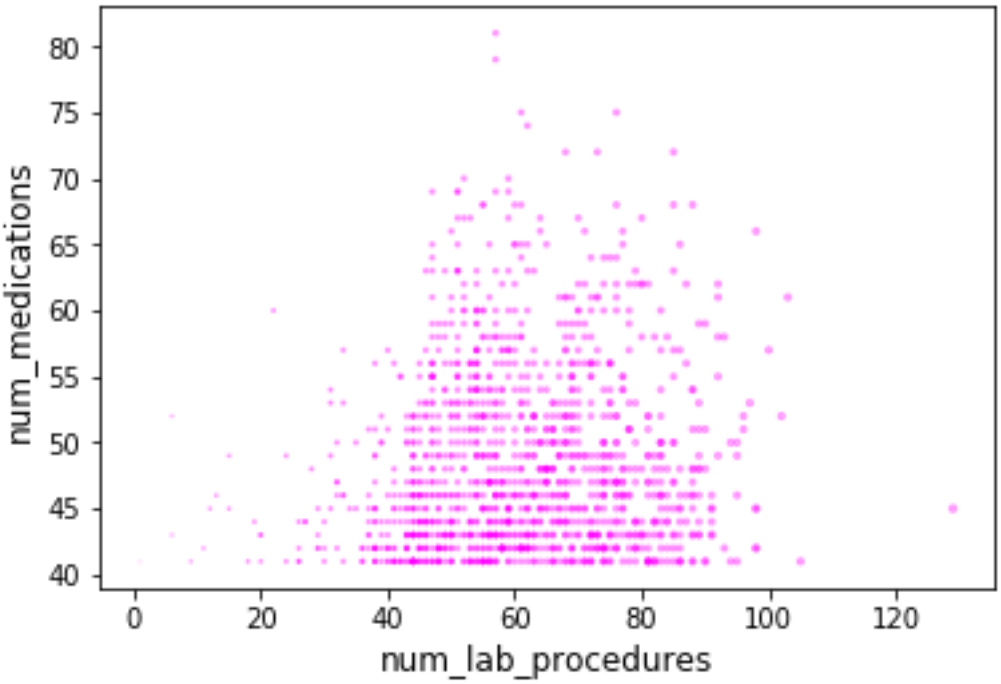
מהגרף הנ״ל, אנו רואים כי קיימת קורולציה אשר מבטאת את הצורך בביצוע מספר רב יותר של בדיקות מעבדה ככל שהמטופל צורך יותר תרופות.

דבר זה נובע מכך שכאשר מטופל צורך מספר רב יחסית של תרופות העולה על התחום הסטנדרטי, דבר זה מצביע על מצב בריאותי מסובך אשר מצריך חקירה ובדיקה מעמיקה יותר אשר מתקיימת בצורת בדיקות מעבדה רבות ומגוונות יותר.

מהגרף ניתן לראות כי רוב אוכלוסיית המטופלים צורכת בין 40-0 תרופות ומספר הבדיקות לא עולה על 80. ניתן לראות כמות מזערית של מטופלים המבצעית מעל 100 בדיקות ולאו דווקא אלו המטופלים שצורכים את כמות התרופות המירבית (ניתן לראות כי המטופלים שעשו מעל 100 בדיקות מעבדה צרכו לא יותר מ50 תרופות).

בנוסף ניתן לראות תופעה מעניינת, קיימת אוכלוסיית מטופלים מסויימת הכוללת 1267 מטופלים אשר מהווים פחות מאחוז מכלל הרשומות, אשר צורכים כמות תרופות במידה גבוהה אשר נעה בין 40 ל 60 תרופות לבן אדם, אשר מתרכזת במספר בדיקות הנע בין 40 ל70.

מניתוח מקרה זה באמצעות טבלת החום המצורפת, המייצגת את הקורלציה בין מספר התרופות אשר הוגשו למטופל לבין מספר בדיקות המעבדה אשר נבדקו למאושפז אפשר לראות כי ישנה התנהגות שונה בין שתי "קבוצות" אשר מיוצגות בצבעים האדומים והכתומים. הקבוצה האדומה מתקבצת באיזור בו שככל שמספר התרופות עולה כך גם מספר בדיקות המעבדה עולות, בעוד שההתנהגות של הקבוה הכתומה שונה לחלוטין ומתפזרת בצורה שונה.

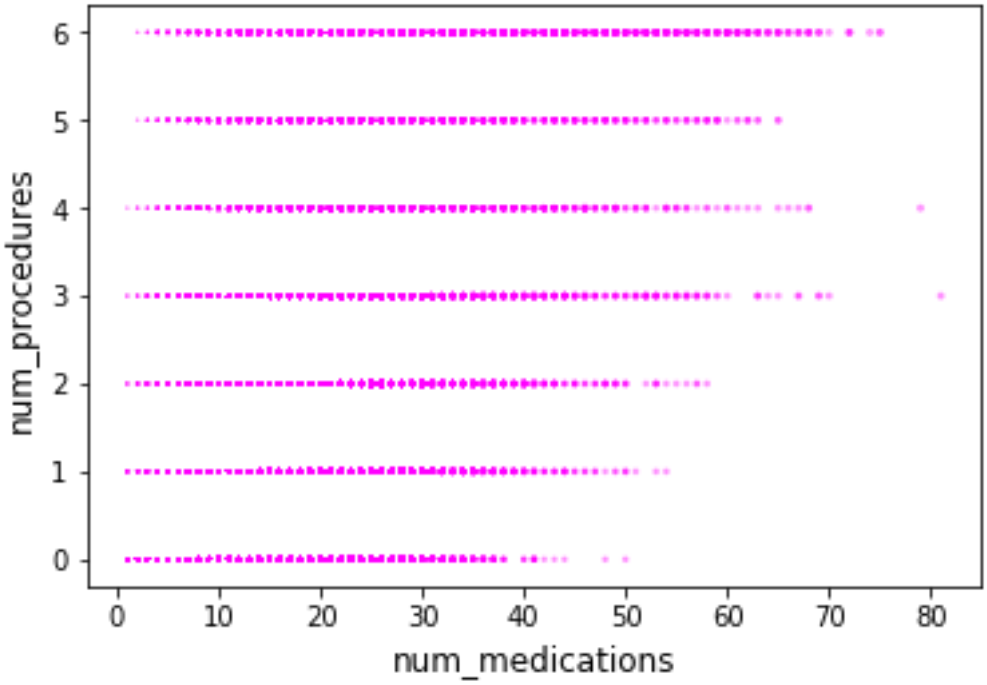


מגרף זה ניתן לראות כי קיימת התפלגות בין מספר התרופות לבין מספר בדיקות המעבדה אשר מתפלגת בצורה נורמלית, כך שהממוצע עומד על 60 בדיקות מעבדה למאושפז, עם סטיית תקן של 15 בדיקות.

עוד ניתן לראות כי 50 % מהמטופלים בקבוצה זו המתבטאת בצבעים הכתומים נבדקים בין 42 ל79 בדיקות מעבדה. אנו רואים כי קיימים מטופלים עם קשרים חריגים בין כמות התרופות שצורך המטופל למספר בדיקות המעבדה שעובר, זאת

ניתן לראות בבירור מהטבלת החום המצורפת. לדוגמא: קיימים 76 מטופלים שעברו 90 בדיקות מעבדה וצורכים מספר דל של 10 תרופות.

Num procedures – num medication



80	70	60	50	40	30	20	10	num_medications
								num_procedures
NaN	NaN	NaN	13	281	4442	23717	15162	0
NaN	NaN	3	44	688	4437	10405	3970	1
NaN	NaN	11	80	651	3274	6013	2014	2
NaN	10	55	135	653	2336	4374	1410	3
1	8	29	126	376	1178	1794	437	4
NaN	9	41	98	300	772	1280	381	5
6	57	176	364	665	1212	1761	422	6

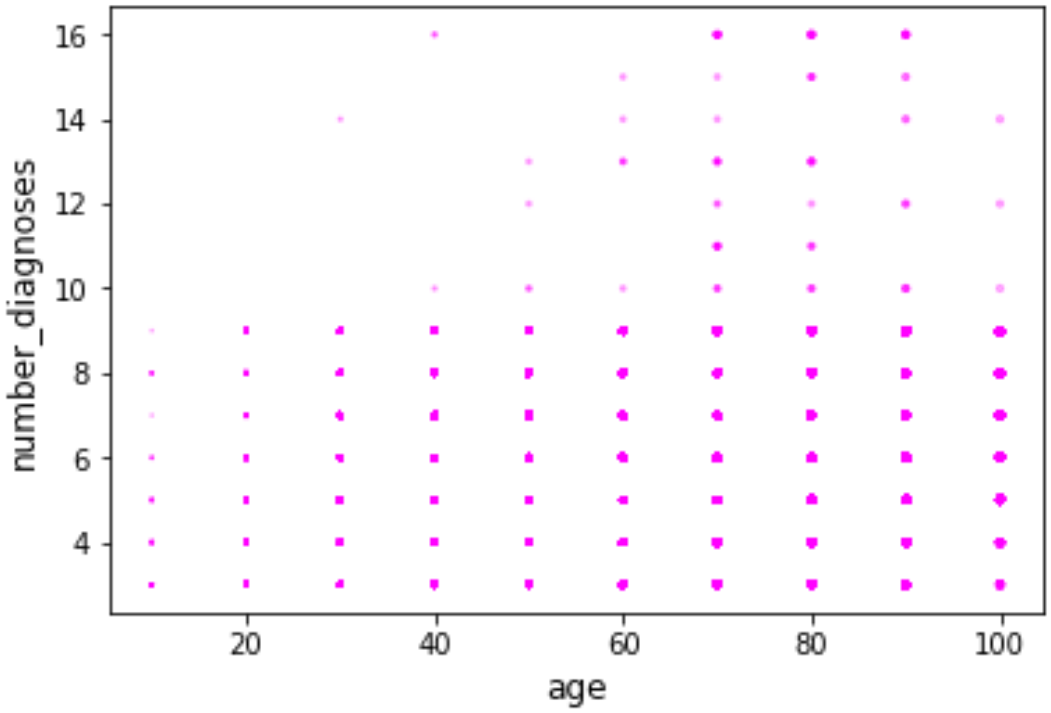
בטבלה הבאה אנו רואים מיפוי של הנתונים בין העמודות של מספר התרופות שקיבל המטופל למספר הטיפולים אשר עבר בבית החולים בזמן האישפוז על ידי מפת חום.

אנו יכולים לראות קשר וקורולציה ברורה בין מספר התרופות אשר קיבל מטופל בעת שהותו בבית החולים למספר הטיפולים אשר קיבל. ככל שמטופל קיבל מספר תרופות מועט יותר כך קיבל מספר טיפולים מועט יותר. אנו רואים את השינוי במידע בצורה ליניארית החל ממספר טיפולים אפסי ועד 10 תרופות לכיוון העליה במספר הטיפולים והתרופות.

הNANים המיוצגים בטבלה מייצגים בינים אשר לא נפלו בהם ערכים מפאת חוסר במקרים קיימים כאלה.

ניתן לראות כי כמות האנשים שצורכים יותר מ50 תרופות מזערית, מה שמראה שישנה העדפה בנתינת טיפולים במקום הגשה מרובה של תרופות המתבטאת בגשה של יותר מ50 תרופות לבן אדם באישפוז.

Number diagnoses – age



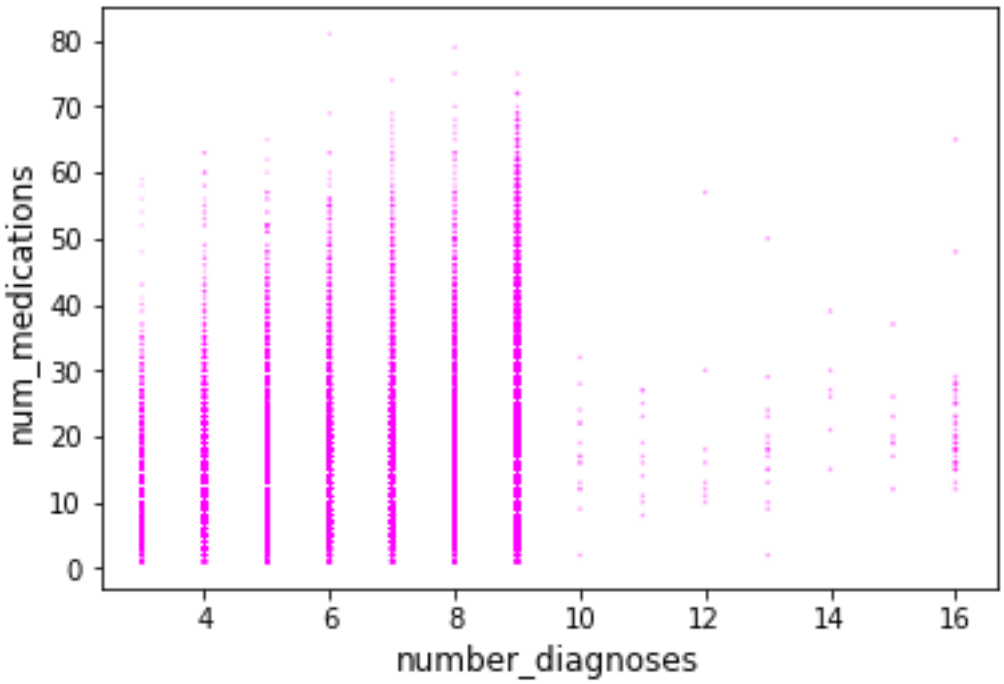
16	15	14	13	12	11	10	9	8	7	6	5	4	3	number_diagnoses
														age
NaN	NaN	NaN	NaN	NaN	NaN	NaN	1	7	1	4	9	13	29	10
NaN	NaN	NaN	NaN	NaN	NaN	NaN	44	24	32	49	92	105	120	20
NaN	NaN	1	NaN	NaN	NaN	NaN	359	120	174	191	279	209	137	30
2	NaN	NaN	NaN	NaN	NaN	1	1125	335	387	484	549	384	270	40
NaN	NaN	NaN	1	1	NaN	2	3608	967	1014	1144	1296	728	428	50
NaN	1	1	3	NaN	NaN	1	7348	1744	1842	1835	2055	1082	592	60
15	1	1	6	2	8	3	10667	2259	2326	2134	2321	1125	526	70
12	4	NaN	6	1	3	3	13038	2618	2542	2400	2445	1081	427	80
9	2	2	NaN	3	NaN	4	9376	1797	1447	1371	1243	512	186	90
NaN	NaN	1	NaN	1	NaN	1	1482	266	239	204	214	83	25	100

בטבלה זו אנו רואים ייצוג קורוליציוני של שנתי העמודות – גיל ביחס לעמודה של מספר האבחנות, על ידי מיפוי חום.

על בסיס הנתונים הללו, אין הרבה להסיק מלבד העובדה שככל שגיל המטופל המאושפז גדל כך גם מספר האבחנות. דבר המעיד על כך שככל שהגיל של המטופל הסוכרתי גדל נוכל לראות על גריעה במצבו הרפואי.

עוד ניתן לראות כי באופן כללי, רוב המטופלים הסוכרתיים מאובחנים ב9 אבחנות, ללא קשר ישיר לגיל, המגיעות באופן ישיר עם מחלת הסוכרת.

Number diagnoses – num medications



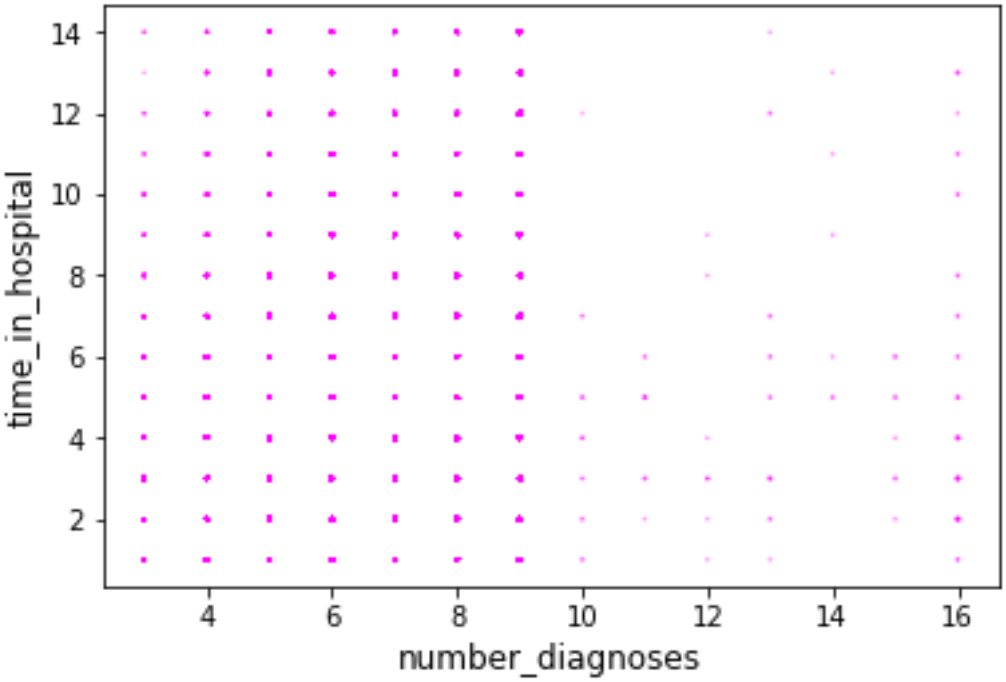
number_diagnoses	3	4	5	6	7	8	9	10	11	12	13	14	15	16
num_medications														
10	1472	2313	4014	3155	2918	2494	7422	2	2	1	3	NaN	NaN	NaN
20	990	2320	5053	5038	5249	5460	25180	8	5	5	9	1	5	21
30	227	549	1152	1266	1435	1744	11245	4	4	1	3	4	2	15
40	42	102	209	249	264	312	2433	1	NaN	NaN	NaN	1	1	NaN
50	4	22	53	76	93	87	523	NaN	NaN	NaN	1	NaN	NaN	1
60	5	14	20	28	33	27	187	NaN	NaN	1	NaN	NaN	NaN	NaN
70	NaN	2	2	3	11	11	54	NaN	NaN	NaN	NaN	NaN	NaN	1
80	NaN	NaN	NaN	NaN	1	2	4	NaN	NaN	NaN	NaN	NaN	NaN	NaN

בטבלה זו אנו רואים ייצוג קורוליציוני של שנתי העמודות – כמות תרופות אשר המטופל קיבל ביחס לעמודה של מספר האבחנות, על ידי מיפוי חום.

גם על בסיס הנתונים הללו, אין הרבה להסיק מלבד העובדה שככל שמספר האבחנות גדל כך מספר התרופות עולה בהתאמה. דבר זה מעיד על העובדה שככל שמספר האבחנות עולה, מצבו הרפואי של המטופל היינו פחות טוב דבר המוביל לצריכת כמות תרופות גבוהה יותר בכדי לטפל באותם אבחנות.

ניתן לראות כי ישנם ערכים "סוררים" אשר לא באים בקו ישר עם ההתנהגות הכללית של המידע בטבלה זו, אשר מתבטאים החל מעשר אבחנות ומעלה והחל מ70 תרופות ומעלה; דבר המוביל אותנו למחיקת ערכים אלה מתוך הבנה כי אלו ערכי קצה אשר לא יתרמו להבנה כללית והסקה מן הדאטא הנ"ל.

Time in hospital – number diagnoses

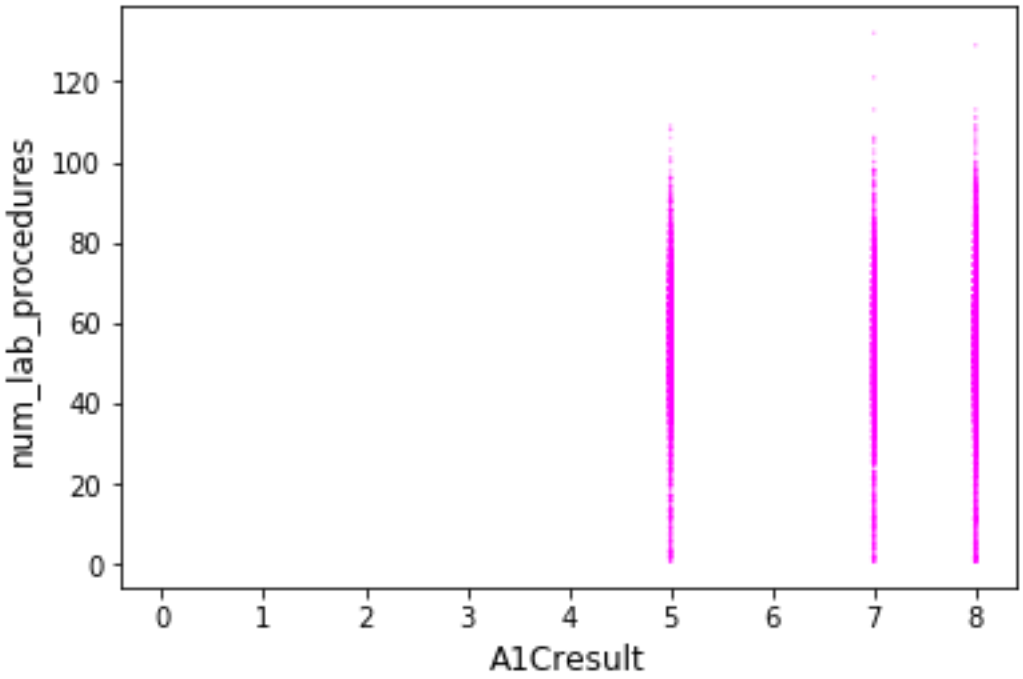


16	15	14	13	12	11	10	9	8	7	6	5	4	3	number_diagnoses
														time_in_hospital
2	NaN	NaN	1	1	NaN	2	4199	1413	1596	1771	2040	1319	771	1
7	1	NaN	2	1	1	2	6517	1710	1957	1873	2069	1249	726	2
7	2	NaN	4	3	3	2	7746	1832	1830	1882	1920	980	521	3
5	1	NaN	NaN	1	NaN	3	6663	1422	1396	1308	1422	673	305	4
3	2	2	2	NaN	5	3	5165	1020	925	856	956	370	175	5
2	2	1	2	NaN	2	NaN	4114	753	655	623	641	252	91	6
2	NaN	NaN	2	NaN	NaN	2	3343	578	503	469	432	157	48	7
2	NaN	NaN	NaN	1	NaN	NaN	2587	400	351	324	336	111	47	8
NaN	NaN	1	NaN	1	NaN	NaN	1797	278	244	202	211	56	16	9
2	NaN	NaN	NaN	NaN	NaN	NaN	1434	239	154	165	149	47	17	10
2	NaN	1	NaN	NaN	NaN	NaN	1144	182	114	121	110	42	8	11
1	NaN	NaN	2	NaN	NaN	1	915	118	118	88	81	28	7	12
3	NaN	1	NaN	NaN	NaN	NaN	747	103	101	71	75	24	2	13
NaN	NaN	NaN	1	NaN	NaN	NaN	677	89	60	63	61	14	6	14

הטבלה הנ״ל מייצגת את היחס הקורוליציוני בין שתי העמודות – עמודת מספר האבחנות למטופל ועמודת מספר הימים אשר שהה המטופל בבית החולים, בעזרת טבלת החום הנ״ל.

קל לראות כי ישנו קשר ישר וקורולציה ברורה בין גובה האבחנות למספר ימי השהיה בבית החולים.

Num lab procedures – A1Cresult



132	130	120	110	100	90	80	70	60	50	40	30	20	10	num_lab_procedures
														A1Cresult
NaN	1	4	17	155	816	3103	8116	12924	17313	15285	8672	5516	7739	0
NaN	NaN	NaN	7	37	157	495	1036	1110	1137	500	141	86	80	5
1	1	1	6	35	111	348	764	873	845	422	122	73	69	7
NaN	1	3	12	92	294	898	1557	1732	1726	826	194	119	100	8

הטבלה הבאה מייצגת את הקשר הקורלציוני בין עמודת מספר בדיקות המעבדה אשר בוצעו על מטופל לבין עמודת ערכי בדיקת רמת ההומוגלובין A1C בדם.

נראה כי ישנו קשר ישיר בין מספר בדיקות המעבדה שבוצעו לקביעת ערך A1C בדם. נסיק כי בכדי לקבוע באופן ברור את גובה המוגלובין A1C בדם נדרשות בממוצע בין 50 ל70 בדיקות מעבדה, כאשר 70 מהווה חסם עליון חלש, מה שמחזק את ירידת ערכי הבדיקות לאחר 70.

סיכום הניתוח הראשוני של הדאטא

הניתוח הראשוני של הדאטא התבצע במספר שלבים ורבדים.

הצגת היסטוגרמות של כלל הפיצ'רים וסינון ראשוני ע"פ המסקנות העולות מההיסטוגרמות הללו. חקירה מעמיקה של ההיסטוגרמות בעלות מידע אינפורמטיבי ובעל פוטנציאל להסקת מסקנות תוך התייחסות לשאלת המחקר בנוגע לחזרת\אי חזרתו של מטופל לאישפוז חוזר בבית החולים.

לאחר מכן חיפשנו מתאמים בין כל שני פיצ'רים וזאת על מנת לזהות פיצרים חשובים. תהליך הקורלציה התבצע במספר שלבים: ראשית השתמשנו במפת חום וטבלת הקורלציות עבור כלל הפיצ'רים של הדאטא. עבור כל זוג פיצ'רים בעלי מתאם מספיק גבוה התבצע תהליך ניתוח מעמיק לשם הסקת מסקנות בנוגע לקשר בין התכונות.

המסקנות העולות מתהליך הניתוח הינן:

נתחיל בהצגת המסקנות אשר עולות מהסתכלות על ההיסטוגרמות.

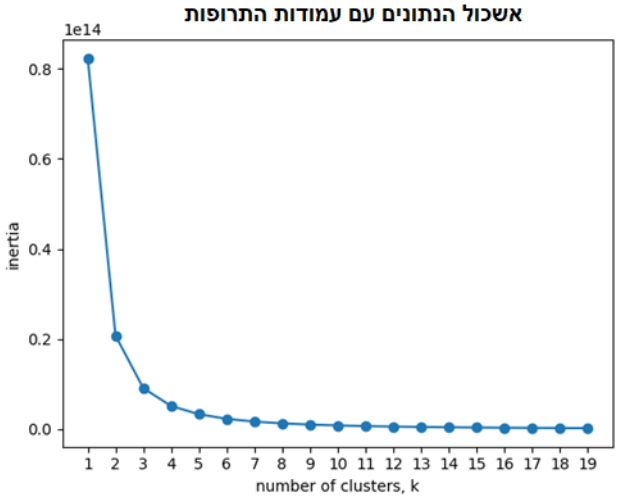
- חלקן הרב של התרופות אינו מתחלק בצורה מגוונת אלא נופל על ערך ספציפי אשר אינו מייחס לנו ערך מחקרי; על כן, מחקנו את עמודות אלה.
- ניתן למחוק את עמודת המין מאחר והיא מתפלגת בצורה אחידה בין הגברים לנשים ולכן לא מספק ערך הסקתי מחקרי, בהתאם לשאלת המחקר שלנו.
- מעמודת הגזע ניתן לראות כי האסייתים מגיבים בצורה הטובה ביותר למחלות סוכרתיות מכלל הגזעים הנבדקים.
- מעמודת הגיל נסיק כי ככל שגילו של אדם עולה כך גם עולים סיכויו לחלות במחלת הסוכרת.
- מעמודת זמן שהות בבית החולים נסיק כי ככל שבן אדם נשאר באישפוז זמן רב יותר כך סיכויו לחזור גודלים יותר.
- מעמודת בדיקות המעבדה נסיק כי ישנן בדיקות ברירת מחדל אשר מהוות בדיקות חובה ראשוניות כאשר מטופל מגיע לאישפוז. כאשר אנו מדברים ספציפית על מחלות סוכרתיות, נראה כי מספר בדיקות המעבדה הממוצעות נעות בין 40 ל60.
- מן עמודות קבלת התרופות נראה כי היא מיוצגת בהתפלגות נורמאלית, כאשר מספר התרופות הממוצע שמטופל צורך עומד על 16 תרופות. נראה כי במקרים נדירים ישנה הגשה של מספר רב מאוד של תרופות המגיע לכדי 70 תרופות למטופל.

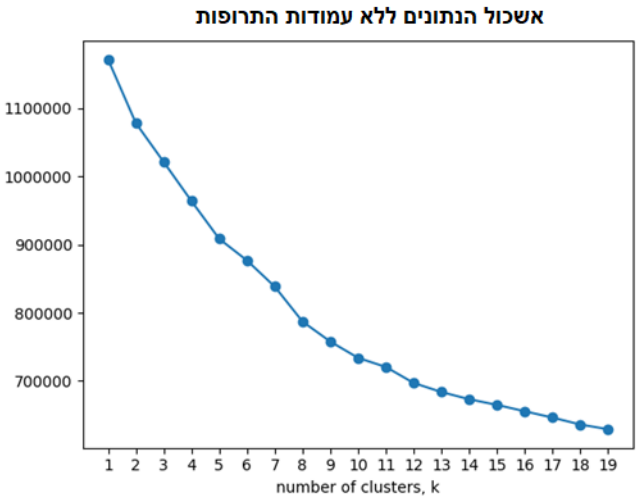
- מעמודת inpatient עולה כי ככל שמטופל חוזר לאישפוז בבית החולים מספר רב יותר של פעמים כך הסיכוי שלו לחזור לאישפוז בטווח זמן קצר עולה.
- מעמודת האבחנות נראה כי מטופלים סוכרתיים מאובחנים ברובם ב9 אבחנות הבאות בקשר ישיר עם מחלת הסוכרת.
- מעמודת בדיקת ההמוגלובין נסיק כי מרבית מן המטופלים אינם נבדקים בבדיקה זו מאחר והיא מהווה בדיקת סף למחלת הסוכרת, דבר המרמז על כך שמרבית מן המטופלים המגיעים כבר מאובחנים מראש כחולים סוכרתיים.
- מן עמודת האינסולין נסיק כי שינוי כלשהו במינון התרופה מוביל לחזרה בסיכוי גבוה בטוח זמנים קצר.

מסקות העולות מתהליך הקורלציה

- בהסתכלות על הקורולציה בין מספר התרופות למספר הימים בבית החולים נסיק כי ככל שמספר ימי השהייה בבית החולים גבוה יותר כך מינון התרופות עולה, כאשר מרבית האנשים צורכים עד 20 תרופות עם הישארות בבית החולים של עד 6 ימים.
- מהקורולציה בין מספר התרופות למספר בדיקות המעבדה נסיק כי ככל שמספר בדיקות המעבדה עולה כך מינון התרופות המגוונות עולה, כאשר אנו מזניחים רשומות חריגות.
- מהקורולציה בין מספר התרופות למספר הבדיקות, ככל שמטופל צורך מספר קטן יותר של תרופות כך מספר הטיפולים קטן.
- מהקורולציה בין עמודת הגיל למספר האחבנות נראה כי ככל שגיל המטופל עולה כך גם מספר האבחנות בצורה ליניארית.
- מהקורולציה בין מספר האבחנות לבין מספר התרופות נראה כי ככל שמספר האבחנות עולה קיימת עליה בכמות התרופות.

בדיקת אופטימליות עבור גודל ה - clusters





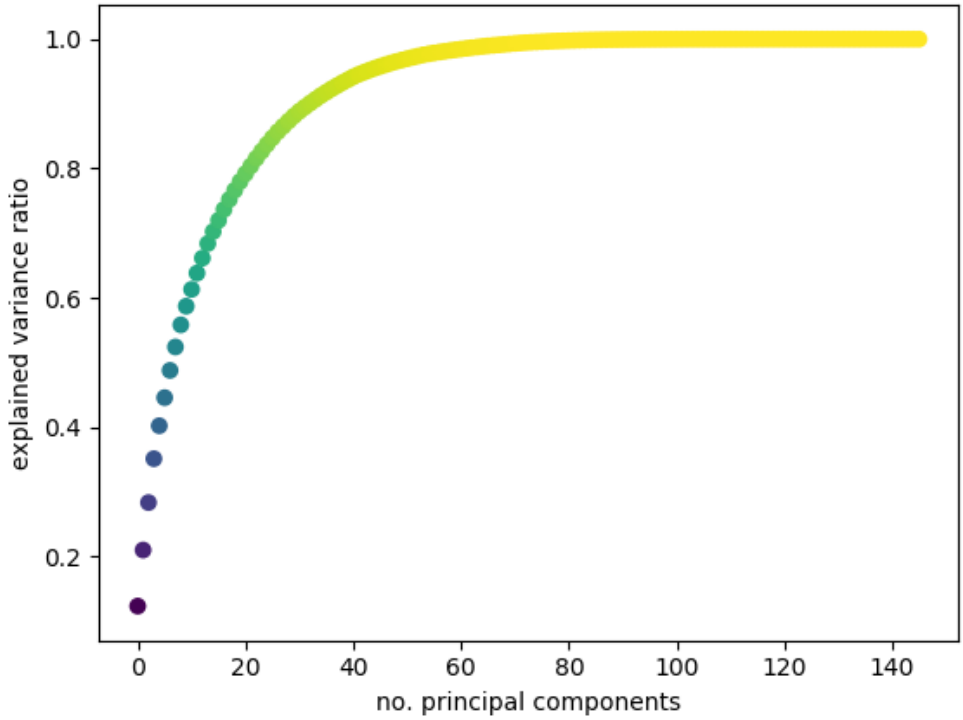
לאחר הטיפול בדאטא, הגענו לשלב ביצוע וחקירת הclustering. התחלנו עם שיטות קלסטריזציה הדורשות כפרמטר את מספר הclusterים אליהם נרצה לחלק את הדאטא.

בכדי למצוא את מספר הקלאסטרים האופטימלי למען הגעה לחלוקה נכונה ומסקנות נבונות והגיוניות השתמשנו באלגוריתם אשר מודד איכות מספר הקלאסטרים ומציג לנו את מספר הקלאסטרים האופטימלי למען חלוקה.

כאשר הפעלנו אלגוריתם זה על שתי סוגי הדאטא שלנו – הדאטא לאחר עיבוד הכולל את כלל התרופות המופיעות בדאטא המקורי ודאטא לאחר עיבוד אשר התרופות מושמטות בו, ראינו בצורה נחרצת כי הדאטא אשר כולל את התרופות מתחלק בצורה אופטימלית לכדי 3 קלאסטרים ראשיים, בעוד שבדאטא המועבד אשר לא כולל את התרופות, אנו רואים התכנסות רק לאחר חלוקת הדאטא ל19 קלאסטרים.

על פי כן, הגענו למסקנה כי נרצה להמשיך לשיטות הורדת מימדים וקלסטריזציה עם הדאטא המעובד הכולל את התרופות כאשר נפעיל את שיטות הקלסטריזציה הדורשות כפרמטר את מספר הקלסטרים אליהם נרצה לחלק את הדאטא עם פרמטר 3.

הורדת מימדים - PCA



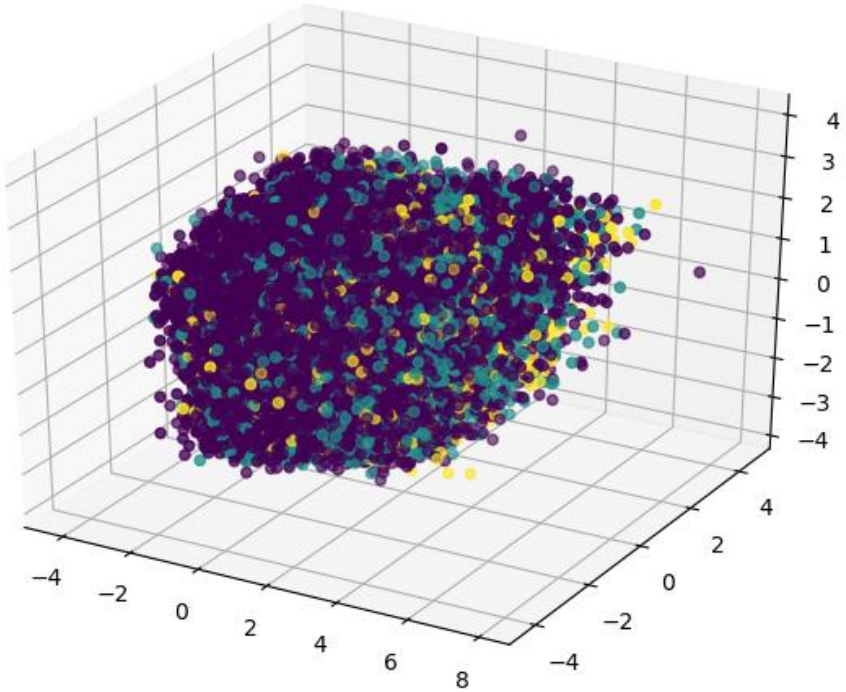
כמו שכתבנו, המשכנו לשיטות הורדת מימדים לאחר בדיקה של חלוקה אופטימלית למספר קלאסטרים בהם הקלסטריזציה מחולקת בצורה הטובה ביותר עם הדאטא המעובד הכולל את עמודות התרופות אשר מכיל 67 אלף רשומות בעיגול גס ועם 146 פיצ'רים.

בכדי למצוא את מספר הפיצ'רים אשר מהווים חלק נכבד מן השונות של הדאטא המעובד, הפעלנו PCA על כל גודל אפשרי של פיצ'רים מ1 עד 146 ומדדנו את רמת השונות שלהם ביחס לכלל הדאטא, כאשר הערכים נעים בין 0 ל1.

הגרף הנ"ל מציג את תוצאות הבדיקה כאשר ניתן לראות שהחל מבסביבות 20 פיצ'רים אנו מקבלים ייצוג נכבד של כלל הדאטא מבחינת שונות.

על פי כן, נמשיך עם דאטא מעובד זה הכולל את התרופות עם הורדת המימדים לכדי 20 פיצ'רים בלבד אל שיטות הקלסטריזציה.

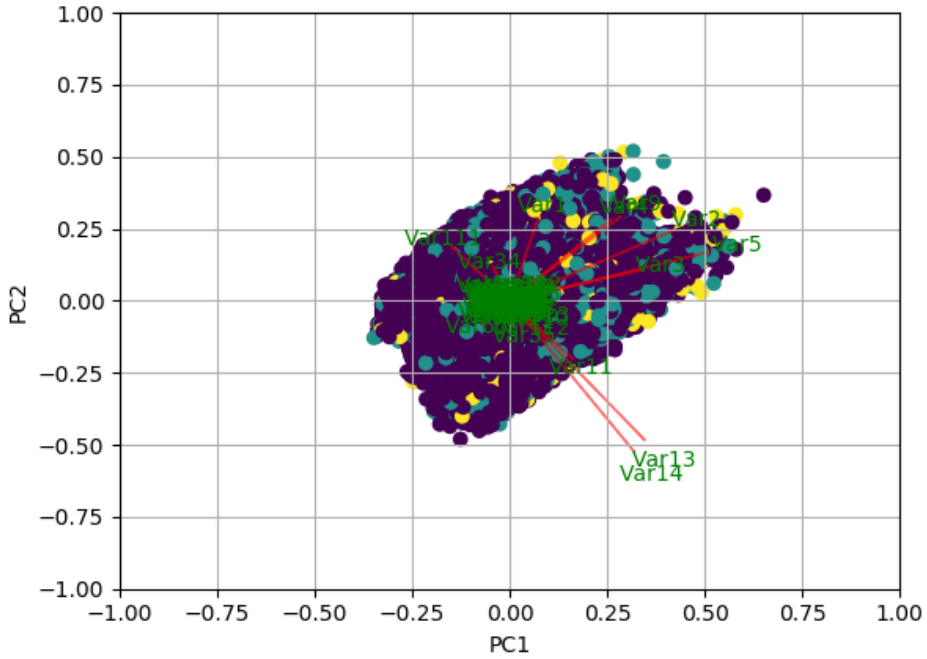
יש לציין שבחנו מספר שיטות הורדת מימדים כגון: KernelPCA, ICA, Factor Anlysis ומצאנו כי PCA מספק לנו מידע בצורה הטובה מבינהם.



הגרף הנ״ל מייצג לנו את הדאטא ללא מחיקת עמודות התרופות, לאחר ביצוע הורדת מימדים בעזרת PCA בתצורת 3 המימדים הדומיננטים ביותר.

בגרף זה הצירים מבטאים את שלושת המימדים הדומיננטים ביותר אשר נוצרו בתהליך הPCA, והצבע מבטא את אחת משלוש אופציות החזרה בעמודה readmitted.

ניתן לראות שאין הסקה חד משמעית בנוגע לחלוקת הקלסטרים ביחס לעמודת הreadmitted.



מוצג להלן מיפוי חשיבות כל פיצ׳ר בעת הורדת המימדים ביחס לערכי העמודה readmitted כאשר הצבעים מייצגים את ערכי העמודה.

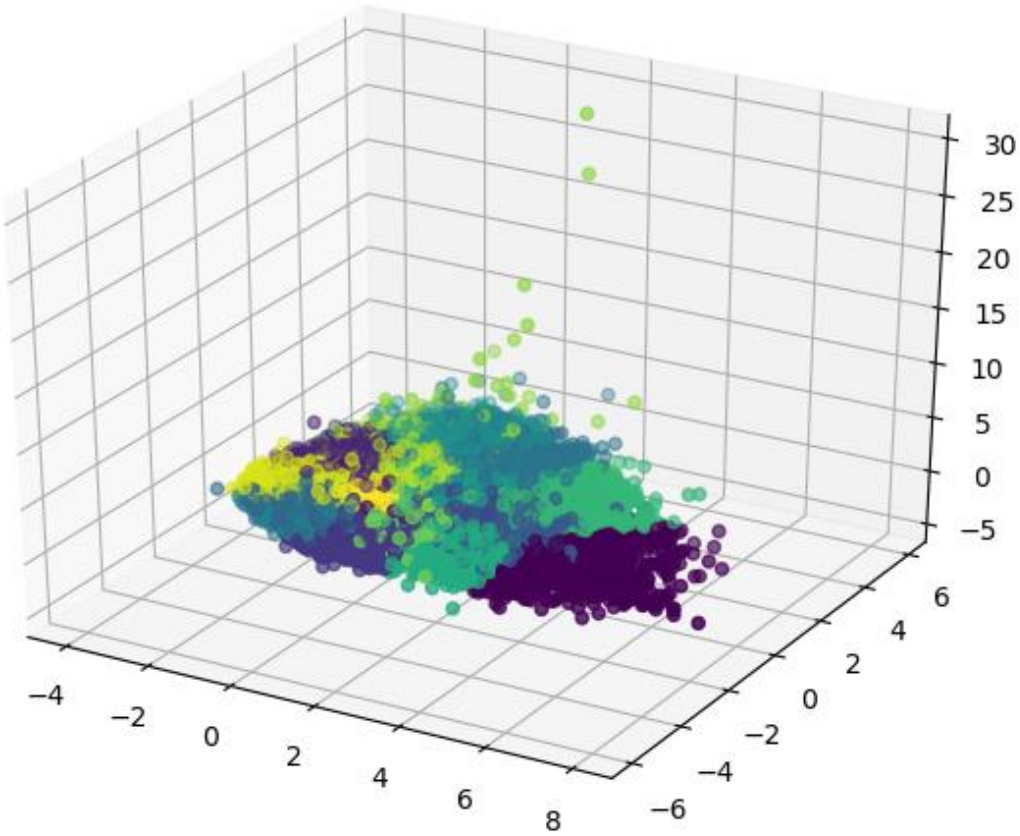
סגול – NO

טורקיז - 30>

צהוב - 30<

הפיצ׳רים החשובים ביותר בעת הורדת המימדים היינם הפיצ׳רים בעלי הmagnitude הגדולים ביותר.

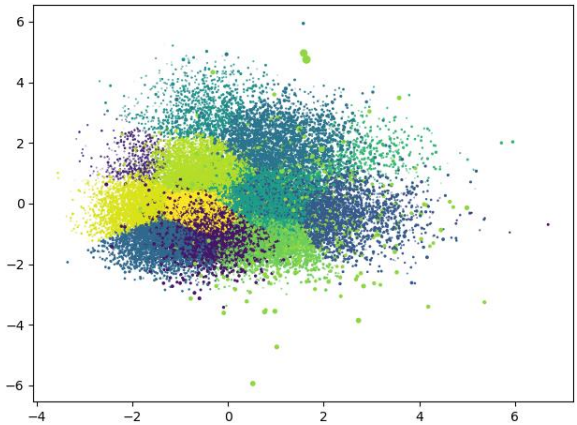
Clustering – עבור הדאטא ללא עמודות התרופות



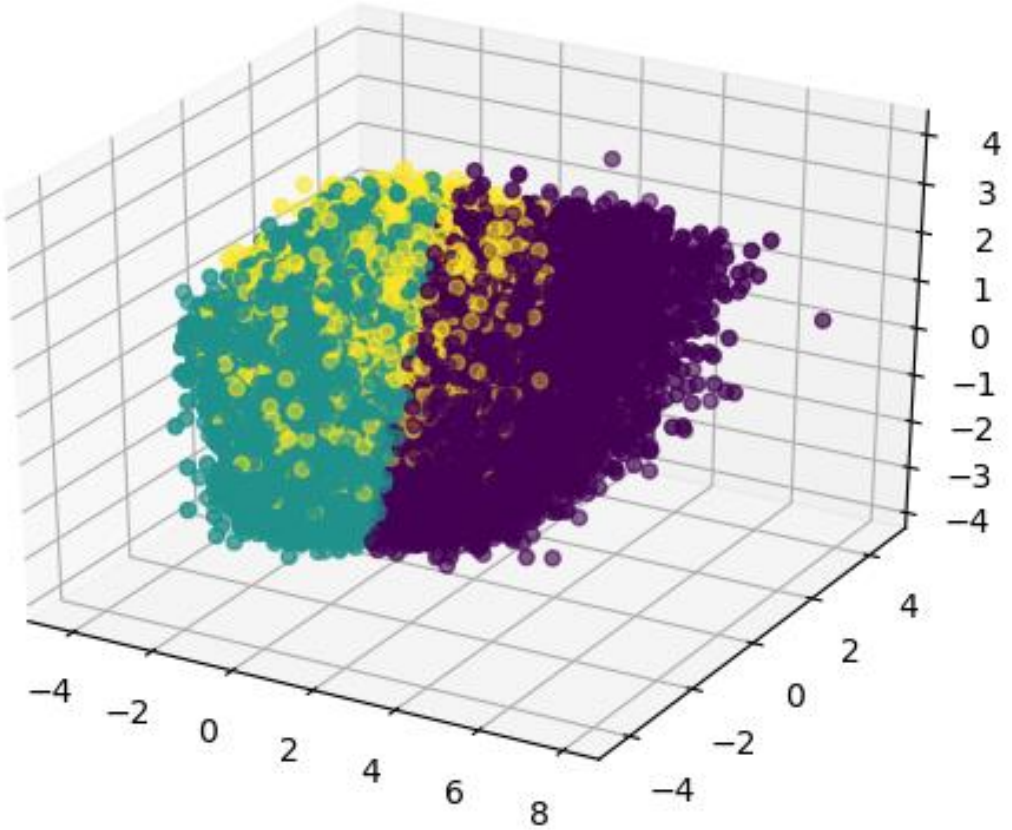
הגרף הנ״ל מייצג לנו את הדאטא ללא מחיקת עמודות התרופות, לאחר ביצוע הורדת מימדים בעזרת PCA בתצורת 3 מימדים דומיננטים ביותר, ולבסוף ביצוע קלסטריזציה על ידי Kmeans כאשר K=19 אשר מהווה מספר קלסטרים אופטימלים כמצויין בסעיפים הקודמים.

אנו רואים כי קיבלנו קלסטרים הניתנים להבחנה, כאשר כל קלאסטר מייוצג בצבע הייחודי לו.

הגרף התחתון מייצג את אותם נתונים כמו הגרף העליון בדו מימד.

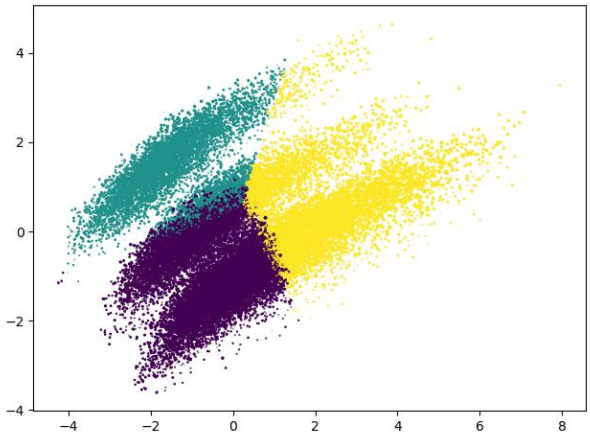


Clustering – עבור הדאטא עם עמודות התרופות



הגרף הנ״ל מייצג לנו את הדאטא לאחר עיבוד ולאחר ביצוע הורדת מימדים בעזרת PCA בתצורת 3 מימדים דומיננטים ביותר, ולבסוף ביצוע קלסטריזציה על ידי Kmeans כאשר K=3 אשר מהווה מספר קלסטרים אופטימלים כמצויין בסעיפים הקודמים.

אנו רואים כי קיבלנו שלושה קלסטרים הניתנים להבחנה, כאשר כל קלאסטר מייוצג בצבע הייחודי לו.



הגרף התחתון מייצג את אותם נתונים כמו הגרף העליון בדו מימד. אנו רואים כי בעצם החלוקה לא מדויקת ולכן נרצה להשתמש בשיטות clustering אחרות שיתאימו ויחלקו את הדאטא שלנו בצורה אופטימלית.

על פי כן, נרצה להשתמש בשיטות קלסטריזציה אשר יחלקו את הדאטא הנ"ל שלנו ל3 קלסטרים מאוזנים כפי שאנו רואים שזה מתחלק.

לכן, בחרנו להשתמש בGGM ובDBscan אשר מטפלים במקרים כאלה כראוי, בכדי שנוכל לקבל את 3 הקלאסטרים הנכונים.

לאחר בדיקת שתי שיטות הקלסטריזציה, מצאנו כי DBscan סיפק לנו את התוצאות הטובות ביותר מבחינת חלוקה ל3 קלאסטרים מופרדים כראוי.